Transactions of Mathematics and Its Applications (2021) **5**, 1–31
https://doi.org/10.1093/imatrm/tnab003

Learning stochastic closures using ensemble Kalman inversion

Tapio Schneider, Andrew M. Stuart and Jin-Long Wu*
California Institute of Technology, Pasadena, CA 91125, USA
*Corresponding author: jinlong@caltech.edu

[Received on 17 April 2020; revised on 30 April 2021; accepted on 28 September 2021]

Although the governing equations of many systems, when derived from first principles, may be viewed as known, it is often too expensive to numerically simulate all the interactions they describe. Therefore, researchers often seek simpler descriptions that describe complex phenomena without numerically resolving all the interacting components. Stochastic differential equations (SDEs) arise naturally as models in this context. The growth in data acquisition, both through experiment and through simulations, provides an opportunity for the systematic derivation of SDE models in many disciplines. However, inconsistencies between SDEs and real data at short time scales often cause problems, when standard statistical methodology is applied to parameter estimation. The incompatibility between SDEs and real data can be addressed by deriving sufficient statistics from the time-series data and learning parameters of SDEs based on these. Here, we study sufficient statistics computed from time averages, an approach that we demonstrate to lead to sufficient statistics on a variety of problems and that has the secondary benefit of obviating the need to match trajectories. Following this approach, we formulate the fitting of SDEs to sufficient statistics from real data as an inverse problem and demonstrate that this inverse problem can be solved by using ensemble Kalman inversion. Furthermore, we create a framework for non-parametric learning of drift and diffusion terms by introducing hierarchical, refinable parameterizations of unknown functions, using Gaussian process regression. We demonstrate the proposed methodology for the fitting of SDE models, first in a simulation study with a noisy Lorenz ’63 model, and then in other applications, including dimension reduction in deterministic chaotic systems arising in the atmospheric sciences, large-scale pattern modeling in climate dynamics and simplified models for key observables arising in molecular dynamics. The results confirm that the proposed methodology provides a robust and systematic approach to fitting SDE models to real data.

Keywords: stochastic differential equation, inverse problem, ensemble Kalman inversion, Gaussian process regression, hierarchical parameterization.

1. Introduction

1.1 *Overview and literature review*

The goal of this paper is to describe a straightforward ensemble-based methodology that facilitates parameter estimation in ergodic stochastic differential equations (SDEs), using statistics derived from time-series data. SDEs arise naturally as models in many disciplines, and the wish to describe complex phenomena without explicitly representing all interacting components within the system makes them of widespread interest. Additionally, the increasing data provide opportunities for the learning of stochastic models in previously unforeseen application domains. However, SDE models, while often accurate at providing statistical predictions, may not be compatible with available data at the small scales where the Itô calculus model asserts very strong almost sure properties rarely found in real data; in particular, the quadratic variation (variance for a pure Brownian motion) is such an almost sure property. For this

reason, standard statistical methodology for parameter estimation in SDEs (see Kutoyants, 2013) is often not suited to fitting models to real data. On the other hand, practical experience in the applied sciences demonstrates the effectiveness of fitting Markovian stochastic processes to sufficient statistics, typically derived from a long time series by averaging procedures.

Ensemble methods have demonstrable success in the solution of inverse problems, based on the use of interacting particle systems driven by the parameter-to-observable map; furthermore, they are robust to noisy evaluations of the map (Duncan *et al.*, 2021). Choosing ergodic averages as observables and viewing finite-time averages, or averages over different initializations, as noisy evaluations of the ergodic average puts us in a setting where we may apply ensemble methods to effectively estimate parameters in SDEs. These ensemble methods are derivative-free, side-stepping thorny technical and computational issues arising in the parameter-to-observable map for SDEs. They are also inherently parallelizable and scale well to the learning of high-dimensional parameter vectors, making them a very attractive computational tool. Finally, the formulation of the inverse problem that we adopt, using time-averaged data, avoids the need to determine the latent trajectory variables; the approach we employ is described in Cleary *et al.* (2021) and Dunbar *et al.* (2021).

Data-driven methods for extracting simplified models are now ubiquitous (Brunton & Kutz, 2019; Coifman *et al.*, 2008; Ferguson *et al.*, 2011; Froyland *et al.*, 2014; Klus *et al.*, 2018; Giannakis, 2019). A guiding example that motivates the need for reduced models of various kinds is the Navier–Stokes equation for fluid motion. It is often too expensive to numerically simulate all relevant degrees of freedom—for example turbulence around an aircraft or convection in Earth's atmosphere. Therefore, stochastic descriptions become important (see, e.g., Majda & Harlim, 2012). In those reduced systems, stochastic models help account for a lack of knowledge and for unresolved variability. Stochastic models are widespread in applications, including in biology (see, e.g., Goel & Richter-Dyn, 2016; Wilkinson, 2018), chemistry (see, e.g., Leimkuhler & Reich, 2004; Tuckerman, 2010; Boninsegna *et al.*, 2018), engineering (see, e.g., Maybeck, 1982), the geophysical sciences (see, e.g., Majda & Kramer, 1999; Palmer, 2001; Arnold *et al.*, 2013) and the social sciences (see, e.g., Diekmann & Mitter, 2014); Gardiner (2009), provides a methodological overview aimed at applications in the physical and social sciences.

The statistics literature concerning parameter estimation from continuous time series often starts from the premise that the diffusion coefficient may be read off from small increments (see, e.g., Kutoyants, 2013), a property that only holds if the data are consistent with an SDE at arbitrarily fine scales. The fact that data may be inconsistent with the SDE at small scales was understood in the finance literature in the early part of this century, and new models were introduced to address this incompatibility (see, e.g., Zhang *et al.*, 2005). In subsequent papers (see, e.g., Pavliotis & Stuart, 2007; Papavasiliou *et al.*, 2009), the setting of multiscale SDEs was used to elucidate similar phenomena arising from models in the physical sciences. Subsequent to these works, new methods were introduced to tackle the inconsistency at small scales, based on subsampling and other multiscale uses of the data (see, e.g., Zhang, 2006; ; Pokern *et al.*, 2009; Abdulle *et al.*, 2021; Callaham *et al.*, 2021; Papaspiliopoulos *et al.*, 2012); see Pavliotis *et al.* (2012) for an overview of some of this work. The potential applicability of problems of this type ranges from applications in econometrics, finance and molecular dynamics, to problems in the geophysical sciences (Cotter & Pavliotis, 2009; Kwasniok & Lohmann, 2009; Ying *et al.*, 2019).

A completely different approach to circumvent the use of fine-scale properties of the time series is to compute sufficient statistics from the time-series and use these to learn parameters. This approach was recently studied as a systematic methodology, and then applied, in a series of papers (see, e.g., Krumscheid *et al.*, 2013, 2015; Kalliadasis *et al.*, 2015), based on statistics computed from multiple realizations and multiple initial conditions; alternatively, one can use a single very long timeseries. In

the community of modeling climate variability, using time averages as data to estimate a stochastic model has been explored for decades (see, e.g., Hasselmann, 1976; Frankignoul & Hasselmann, 1977; Penland & Magorian, 1993). Although most of them focused on linear SDEs, the possible extension to nonlinear SDEs was discussed by Hasselmann (1988). The use of statistics of time series (e.g., moments, autocorrelation functions or power spectral densities) for estimating model parameters is common for discrete time series models, such as autoregressive (AR) or autoregressive moving average (ARMA) models (see, e.g., Brockwell *et al.*, 1991; Neumaier & Schneider, 2001; Lütkepohl, 2013). Another approach to parameter estimation in dynamical systems is known as state augmentation (see, e.g., Anderson, 2001). These methods proceed by augmenting the state to include parameters, and then using state estimation techniques such as the particle filter (see, e.g., Smith, 2013), the unscented Kalman filter (see, e.g., Julier *et al.*, 2000; Albers *et al.*, 2017) and iterative ensemble Kalman smoothers (Evensen, 2019). However, the augmentation approach is notoriously sensitive in many practical settings (see, e.g., Doucet *et al.*, 2001) and we do not pursue it here.

In this paper, we build on the approach pioneered in Krumscheid *et al.* (2013) and use finite-time averaged data. As a consequence, viewing the perfect parameter-to-data map as being an ergodic (infinite-time) average, we have access to only approximate, noisy evaluations of the desired parameter-to-data map, computed from finite-time averages. For this reason, ensemble Kalman methods provide a desirable methodology (Duncan *et al.*, 2021) because they are black-box, derivative-free and robust to noisy parameter-to-data evaluations. A central question when summarizing data, as we do when computing time averages, is whether or not the summary statistics are sufficient to identify the parameters of interest; the reader may refer to the literature on approximate Bayesian computation (ABC) for discussion of this issue (Fearnhead & Prangle, 2012; Sisson *et al.*, 2018). The paper by Wood (2010) provides an example of related ideas applied in the context of learning chaotic dynamical systems; as in our work, and the work by Krumscheid *et al.* (2013), it employs summary statistics that eliminate problems arising from sensitive dependence on initial conditions. Links between ensemble Kalman methods and ABC are discussed in Nott *et al.* (2012).

With the aim of setting our algorithmic approach in context, we now describe the specific form of ensemble Kalman inversion (EKI) that we use in this paper, its relationship to other ensemble Kalman methods that are widely used for both state and parameter estimation and the acronyms used to describe those related algorithms. The methods were originally introduced for state estimation using both filtering (the EnKF, see Evensen, 1994) and smoothing (the ES, see Van Leeuwen & Evensen, 1996). Ensemble methods are now also used to simultaneously estimate the trajectory of a dynamical system, and its parameters. Furthermore, iterating within the 'analysis' step of the data assimilation cycle has proven effective (see Gu & Oliver, 2007; Li & Reynolds, 2009; Bocquet & Sakov, 2012, 2014; Sakov *et al.*, 2012); the terminology ensemble randomized maximum likelihood (sequential-EnRML), iterative ensemble Kalman filter (IEnKF) and iterative ensemble Kalman smoother (IEnKS) is adopted to describe the algorithms used. These approaches have recently been used to learn about model error in dynamical systems, or parameters describing the statistics of model error in dynamical systems (Bocquet *et al.*, 2020; Pulido *et al.*, 2018). However, there are many inverse problems in which dynamics are not present, or in which state estimation is not a desired goal of the computation. The papers (Chen & Oliver, 2012; Emerick & Reynolds, 2013; Evensen, 2018) introduced ensemble Kalman methodologies directly focussed on the solution of general inverse problems, without reference to state estimation in a dynamical system; these methods go by various names, including the iterative ensemble smoother (IES), batch ensemble randomized maximum likelihood (batch-EnRML) and multiple data-assimilation ensemble smoother (ES–MDA). In the context of solving inverse problems, the methods IES, batch-EnRML and ES–MDA adopt a Bayesian framework and iterate over a prescribed set of iterations,

starting from prior samples, with goal being the generation of approximate posterior samples. The seminal paper by Reich (2011) made an important step by connecting ensemble methods with sequential Monte Carlo for Bayesian inverse problems and additionally remarks that ensemble Kalman methods could be used for for optimization (classical rather than Bayesian inversion) and discusses possible stopping criteria.

The ensemble Kalman methods are formulated and evaluated as an optimization approach to general inverse problems in Iglesias *et al.* (2013) and the idea of incorporating constraints, widely useful in applications, is described in Albers *et al.* (2019). Iglesias *et al.* (2013) demonstrate that, for a variety of PDE inverse problems, ensemble Kalman methods implemented without localization perform well. These methods may be viewed as minimizing the model-data misfit, and regularization is introduced through the invariant subspace property of the algorithm: the minimization is confined to the space spanned by the initial ensemble. We refer to this methodology, and its continuous time analogues (see Schillings & Stuart, 2017), as EKI, recognizing that it is a variant on the creative ideas developed in Chen & Oliver (2012) and Emerick & Reynolds (2013). We emphasize, however, that the goal of the EKI approach is to solve an optimization formulation of the inverse problem, rather than the Bayesian inverse problem that motivates the approaches in Chen & Oliver (2012) and Emerick & Reynolds (2013). Overfitting can still be an issue for EKI methods and may be addressed by generalizing ideas standard in the classical optimization approach to inverse problems (Engl *et al.*, 1996), as anticipated in Reich (2011); in particular, Levenberg–Marquadt analogues of EKI are studied in Iglesias (2015, 2016), and a Tikhonov extension, TEKI, is described in Chada *et al.* (2020). Bayesian regularization is also possible, leading to the ensemble Kalman sampler in Garbuno-Inigo *et al.* (2020b). In this paper, the problems studied typically have free parameters with dimension smaller than or similar to the dimension of the data-set; thus overfitting is not an issue and regularization is not employed. Other derivative-free optimization methods could also be employed, such as the consensus-based optimization procedures described in Carrillo *et al.* (2018). An important aspect of the success of the ensemble Kalman methods we use is their affine invariance, a concept introduced in Goodman & Weare (2010) for Monte Carlo methods; its significance for ensemble methods was identified in Garbuno-Inigo *et al.* (2020a) and explains the problem-independent convergence rates obtained by the method, provably in the case of linear problems Garbuno-Inigo *et al.* (2020b).

Within the EKI-based parameter estimation methodology, we employ ideas from Gaussian process regression (GPR) (see, e.g., Rasmussen & Williams, 2006) to parameterize unknown functions; this refinable, hierarchical approach to function representation leads to novel nonparametric methods for function learning. Our approach builds on preceding, nonhierarchical approaches to inversion using GPR such as that described in Xiao *et al.* (2016). The concept of learning the values at some fixed nodes of a Gaussian process, known as the pilot point method (see, e.g., Doherty *et al.*, 2010), has also been extensively explored by the groundwater modeling community in the context of inverse problems.

1.2 *Our contributions*

Our contributions in this paper are as follows:

1. We formulate parameter estimation in ergodic SDEs as a classical inverse problem for the parameter-to-data map defined by ergodic averaging.
2. We develop algorithms suited to the setting in which the parameter-to-data map is available only through finite-time averages, which may be viewed as providing noisy approximations of the ideal ergodic averages.

3. Through a sequence of examples described below, we demonstrate that a simple and straightforward implementation of ensemble Kalman methods, the EKI, is well-adapted to solving the inverse problem in which only noisy approximate evaluations of the parameter-to-data map are available.
4. Within the EKI we demonstrate the utility of hierarchical parameterizations of unknown functions, using GPR.

We demonstrate the methodology when applied to a variety of examples:

- A simulation study that employs a noisy (SDE) version of the Lorenz '63 model.
- Reduction of the Lorenz '63 ODE model to a two-dimensional SDE in coordinates computed by applying PCA to the ODE data.
- The multiscale Lorenz '96 ODE model, seeking an SDE closure in the slow variables alone.
- To fit a stochastic delay differential equation (SDDE) to El Niño–Southern Oscillation data.
- To fit an SDE that describes fluctuations in the dihedral angle of a butane molecule.

In considering these simulation studies and real-data examples, we demonstrate the effectiveness of the methodology to find parameters, and we evaluate the accuracy of various stochastic models derived from data. In section 2, we formulate the inverse problem of interest and introduce four example problems to which we will apply our methodology. Section 3 describes the ensemble Kalman methodology we employ to solve the inverse problem, as well as a discussion of the novel hierarchical Gaussian process based representation that we employ to represent, and learn, unknown functions. In section 4, we describe numerical results relating to each of the four example problems. We conclude in section 5.

2. Problem formulation

The aim of this work is to estimate parameters $\theta \in \Theta$ in the SDE

$$\frac{dx}{dt} = f(x;\theta) + \sqrt{\Sigma(x;\theta)}\frac{dW}{dt} \tag{2.1}$$

where $x \in \mathbb{R}^n, f : \mathbb{R}^n \times \Theta \mapsto \mathbb{R}^n$ and $\Sigma : \mathbb{R}^n \times \Theta \mapsto \mathbb{R}^{n\times n}$. For all of the examples $\Theta \subseteq \mathbb{R}^p$, but the algorithms we use extend to include the nonparametric setting in which $p = \infty$. However, in the specific applications considered here, p is small, of $\mathcal{O}(10)$; and in many other applications envisaged, p may be considerably smaller than the dimension n of the state space. Among several parameterizations used in this paper, we showcase a GPR based method for hierarchical function representation that is refinable.

We assume that the SDE is ergodic and let $\mathbb{E}$ denote expectation with respect to the stationary process resulting from this ergodicity. If $x(\cdot;\theta) \in \mathcal{X} := C(\mathbb{R}^+;\mathbb{R}^n)$ denotes a solution of the SDE started in a statistical stationary state and $\mathcal{F} : \mathcal{X} \mapsto \mathbb{R}^q$ is a function on the space of solution trajectories, where q denotes the dimension of data space, then define $\mathcal{G} : \Theta \mapsto \mathbb{R}^q$ by

$$\mathcal{G}(\theta) = \mathbb{E}\mathcal{F}(x(\cdot;\theta)).$$

We wish to solve the inverse problem of finding θ from noisy approximate evaluations of $\mathcal{G}(\theta)$. The noisy approximate evaluations arise from the fact that we will use finite-length trajectories to approximate the expectation defining $\mathcal{G}(\theta)$; averages over initial conditions and/or realizations of the noise could also be used. The following remark highlights the various observables $\mathcal{F}$ that we will use in this paper.

REMARK 1 We will use $m^{\text{th}}-$moments of vector x at time $t = 0$:

$$\mathcal{F}_m(x(\cdot)) = \Pi_{j\in M} x_j(0),$$

where x_j denotes the j^{th} element of vector x, and M is a subset of cardinality m comprising indices (repetition allowed) from $\{1, \cdots, n\}$, leading to the ergodic average $\mathcal{G}_m$. We will also use $\mathcal{F}_{ac}(x(\cdot)) = x(t) \otimes x(0)$, leading through ergodic averaging to the auto-correlation function $\mathcal{G}_{ac}$ of the stationary process. And finally we will use $\mathcal{G}_{psd}$ to denote parameters of a polynomial fit to the logarithm of the power spectral density (PSD); recall that the PSD is the Fourier transform of the auto-correlation function of the stationary process. All of the functions $\mathcal{G}_m, \mathcal{G}_{ac}$ and $\mathcal{G}_{psd}$ can be approximated by time-averaging.

It is instructive to think of $\mathcal{G}(\theta)$ as the infinite-time average of the quantities of interest so that, assuming ergodicity, the dependence of the initial condition of the trajectory generating the data disappears. By doing so, we obtain an inverse problem in which the latent variable, the trajectory itself, disappears from the inference problem. This is distinct from many other approaches to parameter estimation in which trajectories and parameters are jointly inferred (Bocquet *et al.*, 2020; Pulido *et al.*, 2018). In practice, we have only finite time averages available, which means that $\mathcal{G}(\theta)$ is only available to us through approximate, noisy evaluations, with noise entering through the dependence on the initial condition and through sensitive dependence on initial conditions. The ensemble methods described in the next section are demonstrably and provably effective in dealing with the setting in which only approximate, noisy evaluations of $\mathcal{G}(\theta)$ are available (Duncan *et al.*, 2021).

We now describe four examples that will be used to illustrate the methodology.

EXAMPLE 1 (Lorenz 63 System). The Lorenz equations (see, e.g., Lorenz, 1963) are a system of three ordinary differential equations taking the form

$$\dot{x} = f(x), \tag{2.2}$$

where $x = [x_1, x_2, x_3]^\top$ and $f : \mathbb{R}^3 \mapsto \mathbb{R}^3$ is given by

$$\begin{aligned} f_1(x) &= \alpha(x_2 - x_1), \\ f_2(x) &= x_1(\rho - x_3) - x_2, \\ f_3(x) &= x_1 x_2 - \beta x_3. \end{aligned} \tag{2.3}$$

We are interested in the noisy version of these equations, in the form

$$\dot{x} = f(x) + \sqrt{\sigma}\dot{W}. \tag{2.4}$$

This SDE will be used in a simulation study to illustrate our methodology.

We will also use the Lorenz 63 model (2.2), (2.3) written in a new coordinate system computed by means of PCA, as introduced in Selten (1995) and Palmer (2001). This amounts to introducing new coordinates $a = Ax$, in which we obtain

$$\dot{a} = g(a), \tag{2.5}$$

where $x = [x_1, x_2, x_3]^\top$ and $g : \mathbb{R}^3 \mapsto \mathbb{R}^3$ is given by

$$\begin{aligned} g_1(a) &= 2.3a_1 - 6.2a_3 - 0.49a_1a_2 - 0.57a_2a_3, \\ g_2(a) &= -62 - 2.7a_2 + 0.49a_1^2 - 0.49a_3^2 + 0.14a_1a_3, \\ g_3(a) &= -0.63a_1 - 13a_3 + 0.43a_1a_2 + 0.49a_2a_3. \end{aligned} \tag{2.6}$$

The coordinate a_3 contains around 4% of the total variance of the system. In Palmer (2001), this was used as an argument to seek a stochastic dimension reduction of the model, in discrete time, in the variables a_1, a_2 alone. We reinterpret this idea in continuous time and use our methodology to evaluate the idea, using data from (2.5), (2.6) to study the fidelity possible when fitting an SDE of the form

$$\begin{aligned} \dot{a}_1 &= 2.3a_1 - 0.49a_1a_2 + \psi_1(a_2) + \sqrt{\sigma_1(a_2)}\dot{W}, \\ \dot{a}_2 &= -62 - 2.7a_2 + 0.49a_1^2 + \psi_2(a_1) + \sqrt{\sigma_2(a_1)}\dot{W}, \end{aligned} \tag{2.7}$$

where the functions $\psi_1(\cdot)$, $\psi_2(\cdot)$, $\sigma_1(\cdot)$, $\sigma_2(\cdot)$ are represented by GPR as described in detail subsection 3.2.

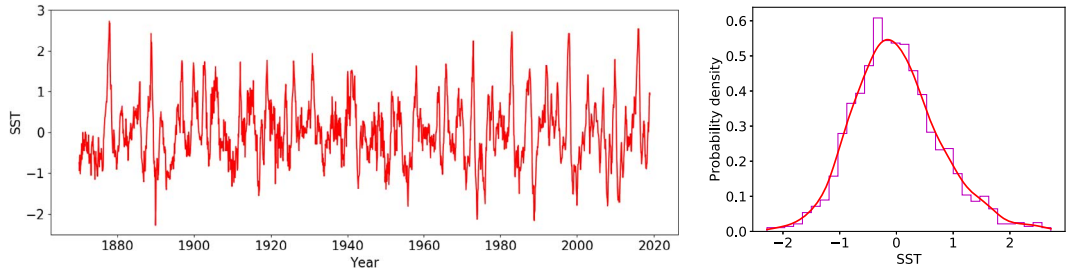
EXAMPLE 2 (Lorenz 96 System). The Lorenz 96 multiscale system (see, e.g., Lorenz, 1996) describes the evolution of two sets of variables, denoted by x_k (slow variables) and $y_{j,k}$ (fast variables):

$$\begin{aligned} \frac{dx_k}{dt} &= -x_{k-1}(x_{k-2} - x_{k+1}) - x_k + F - hc\overline{y}_k, \quad k \in \{1, \dots, K\}, \\ \frac{1}{c}\frac{dy_{j,k}}{dt} &= -by_{j+1,k}(y_{j+2,k} - y_{j-1,k}) - y_{j,k} + \frac{h}{J}x_k, \quad (j,k) \in \{1, \dots, J\} \times \{1, \dots, K\} \\ x_{k+K} &= x_k, \quad y_{j,k+K} = y_{j,k}, \quad y_{j+J,k} = y_{j,k+1}. \end{aligned} \tag{2.8}$$

The coupling term $hc\overline{y_k}$ describes the impact of fast dynamics on the slow dynamics, with $\overline{y}_k$ being the average

$$\overline{y}_k = \frac{1}{J}\sum_{j=1}^{J} y_{j,k}. \tag{2.9}$$

We work with the parameter choices as in Schneider *et al.* (2017), which are the same parameter choices as in Lorenz (1996). Specifically, we choose $K = 36$, $J = 10$, $h = 1$ and $F = c = b = 10$. By assuming a spatially homogeneous (with respect to j) equilibrium in the fast dynamics, fixing k and x_k as in



(a) Time series (b) Histogram

FIG. 1. Illustration of ENSO data (Rayner *et al.*, 2003) from year 1870 to 2019 (the time interval between two adjacent data points is one month), where SST stands for sea surface temperature. It should be noted that although the left-hand shows a time series, only time-averaged statistics are used as training data in this work.

Fatkullin & Vanden-Eijnden (2004), we obtain

$$\bar{y}_k = \frac{h}{J} x_k. \tag{2.10}$$

We will seek a stochastic closure for the slow variables $\{x_k\}$ encapsulating systematic deviations from, and random fluctuations around, this simple balance:

$$\begin{aligned} \dot{X}_k &= -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F - \frac{h^2 c}{J} X_k + \psi(X_k) + \sqrt{\sigma}\dot{W}, \\ X_{k+K} &= X_k. \end{aligned} \tag{2.11}$$

Using data from (2.8), we will fit a parameterized $\psi(\cdot)$ and the noise level σ.

EXAMPLE 3 (El Niño–Southern Oscillation). The El Niño–Southern Oscillation (ENSO) (see, e.g., Rayner *et al.*, 2003) is a well-documented phenomenon, which describes irregularly recurring changes in central and eastern tropical Pacific Ocean temperatures. Figure 1 (time series and histogram) shows Pacific sea surface temperature (SST) data (mean value has been removed) for years 1870 to 2019. The temperature data are averaged within 5S-5N and 170-120W, a region known as Niño 3.4. It is postulated that the mechanism of ENSO can be illustrated by a delayed oscillator model with two time delays determined by properties of the dynamics of the central tropical Pacific Ocean. Specifically, the two time delays are often interpreted as associated with an eastward traveling Kelvin wave and a westward traveling Rossby wave (which also becomes an eastward traveling Kelvin wave after being reflected by the western coastline) (see, e.g., Tziperman *et al.*, 1994). Thus, the time delays can be viewed as known and estimated quantitatively from the corresponding wave speeds.

The aim is to fit a SDDE (see, e.g., Buckwar, 2000; Erneux, 2009) to the ENSO data shown in Fig. 1 (time-series and histogram). To be concrete, we will fit a model of the following form (Tziperman *et al.*, 1998):

$$\frac{dx}{dt}(t) = a\tanh(x(t-\tau_1)) - b\tanh(x(t-\tau_2)) - cx(t) + \sqrt{\sigma}\frac{dW}{dt}(t). \tag{2.12}$$

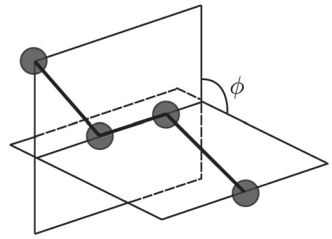
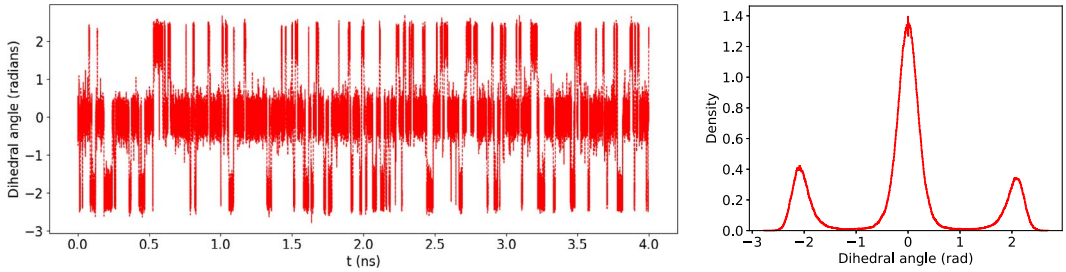


FIG. 2. The definition of butane molecule dihedral angle.



(a) Time series (b) Histogram

FIG. 3. Illustration of the true butane dihedral angle data (the time interval between two adjacent data points is 10^{-6} ns).

Here delay τ_1 represents the effect of the eastward traveling Kelvin wave, and delay τ_2 represents the effect of the westward traveling Rossby wave. Since good estimates for τ_1 (1.15 months) and τ_2 (5.75 months) exist, we will simply fit the four parameters a, b, c, σ to time-averaged data computed from the time-series shown in the left-hand panel in Fig. 1.

EXAMPLE 4 (Butane molecule dihedral angle). In many problems arising in molecular dynamics, it is of interest to determine reduced models describing the behavior of certain functionals derived from the molecular conformation. An example of such a functional is the dihedral angle in a simple model of a butane molecule (see, e.g., Schlick, 2010). Figure 2 shows how the dihedral angle is defined from the four carbon atoms that comprise the butane molecule. Figure 3 shows the time series and histogram for this angle, derived from a molecular dynamics model that we now describe. The model comprises a system of 12 second-order SDEs for the four atomic positions in $\mathbb{R}^3$, of Langevin type:

$$m_0 \frac{d^2x}{dt^2} + \gamma_0 \frac{dx}{dt} + \nabla V(x) = \sqrt{\frac{2\gamma_0}{\beta_0}} \frac{dW}{dt}. \tag{2.13}$$

The potential V, mass m_0, damping γ_0 and inverse temperature β_0 characterize the molecule. From the time series of $x \in C(\mathbb{R}^+; \mathbb{R}^{12})$ generated by this model we fit a simplified model for the the dihedral angle. This simplified model for $\phi \in C(\mathbb{R}^+; \mathbb{R})$ takes the form

$$\frac{d^2\phi}{dt^2} + \gamma(\phi) \frac{d\phi}{dt} + \nabla \Psi(\phi) = \sqrt{2\sigma\gamma(\phi)} \frac{dW}{dt}. \tag{2.14}$$

We will fit parameterized versions of γ and Ψ to data for ϕ generated by studying the time series for the dihedral angle defined by x from (2.13). Related work, fitting SDE models for the dihedral angle, may be found in Papaspiliopoulos *et al.* (2012) and Pokern *et al.* (2009). For a broader introduction to the subject of finding Markov models for bimolecular dynamics see Djurdjevac *et al.* (2010); Ferguson *et al.* (2011); Schütte & Sarich (2013); Zhang *et al.* (2017).

3. Algorithms

In subsection 3.1, we describe the ensemble-based derivative-free optimization method that we use to fit parameters. The approach is based on the algorithms pioneered in Chen & Oliver (2012) and Emerick & Reynolds (2013) but is aimed at solving an optimization formulation of the parameter estimation problem, rather than a Bayesian formulation; we refer to the specific iterated form of the algorithm used here as an EKI algorithm. In subsection 3.2, we discuss how we use GPR to design hierarchical, refinable parameterizations of unknown functions that we wish to learn from data.

3.1 *Ensemble Kalman inversion*

Recall that we view the data that we are given, $y \in \mathbb{R}^J$, as a noisy evaluation of the function $\mathcal{G}(\theta)$ defined by ergodic averages. We use the notation $G_\tau(\theta; x_0)$ to denote this noisy evaluation, which arises from finite-time averaging of duration τ, started at initial condition x_0; this finite-time averaging approximates the ergodic average ($\tau = \infty$) in which dependence on x_0 disappears. Appealing to central limit theorem results that quantify rates of convergence towards ergodic averages, the inverse problem can be formulated as follows: given $y \in \mathbb{R}^J$, find $\theta \in \Theta$ so that

$$y = G_\tau(\theta; x_0) \approx \mathcal{G}(\theta) + \eta, \quad \eta \sim N\big(0, \Gamma(\theta)\big). \tag{3.1}$$

Although this central limit theorem argument leads to a Γ which is $\theta-$dependent, in practice we make the approximation that it is constant and solve the resulting Bayesian inverse problem

$$y = \mathcal{G}(\theta) + \eta, \quad \eta \sim N\big(0, \Gamma\big). \tag{3.2}$$

In this formulation, the trajectory initial condition x_0 disappears from the inference problem. However, $\mathcal{G}(\theta)$ is not computable and only available to us through noisy approximate evaluations. We address this issue after introducing the EKI algorithm below. Estimating Γ may be achieved by using time-series of different lengths, as explained in Cleary *et al.* (2021) in the specific case of $\mathcal{G}$ derived from moments, $\mathcal{F}_m$.

The natural objective function associated with the inverse problem (3.2) is

$$\frac{1}{2}\left\| \Gamma^{-\frac{1}{2}} \big(y - \mathcal{G}(\theta)\big) \right\|^2. \tag{3.3}$$

The two primary reasons for using EKI to find approximate minimizers of this objective function are the following: (a) it does not require derivatives, which can be difficult to compute for SDEs; (b) it is robust to noisy evaluations of the forward map. The method behaves (provably in the linear case, approximately in the nonlinear case) like a projected gradient descent (Schillings & Stuart, 2017; Garbuno-Inigo *et al.*, 2020b), despite not computing derivatives, and the use of differences promotes a desirable averaging making it robust to noisy evaluations of the forward map (Duncan *et al.*, 2021).

The EKI algorithm we employ is described in Iglesias *et al.* (2013) and Albers *et al.* (2019). The algorithm propagates a set of J parameter estimates $\{\theta_n^{(j)}\}_{j=1}^J$ through algorithmic time n, using the update formula

$$\theta_{n+1}^{(j)} = \theta_n^{(j)} + C_n^{\theta G}\left(C_n^{GG} + \Gamma\right)^{-1}\left(y - \mathcal{G}(\theta_n^{(j)})\right). \tag{3.4}$$

The matrix C_n^{GG} is the empirical covariance of $\{\mathcal{G}(\theta_n^{(j)})\}_{j=1}^J$, while matrix $C_n^{\theta G}$ is the empirical cross-covariance of $\{\theta_n^{(j)}\}_{j=1}^J$ with $\{\mathcal{G}(\theta_n^{(j)})\}_{j=1}^J$.

We implement the preceding algorithm with two modifications. The first reflects the benefits that can accrue from randomizing the data, helping to expand the search through parameter space, analogous to stochastic gradient descent (Goodfellow *et al.*, 2016). The second reflects the fact that the forward model $\mathcal{G}$ represents ergodic averages, but is only available to us approximately through finite time averages which depend on initial condition. Thus, the algorithm we implement has form

$$\theta_{n+1}^{(j)} = \theta_n^{(j)} + C_n^{\theta G}\left(C_n^{GG} + \Gamma\right)^{-1}\left(y_n^{(j)} - G_n^{(j)})\right). \tag{3.5}$$

We now detail exactly how $y_n^{(j)}$ and $G_n^{(j)}$ are defined. In the experiments reported in this paper, we add i.i.d. (w.r.t. j and n) random mean zero Gaussian noise with covariance Γ to the data y to obtain $y_n^{(j)}$; however, we have verified that near identical parameter estimates are obtained in the same number of iterations, without adding any random noise to y, for several of the experiments reported. Randomization should be interpreted in the same way that stochastic gradient descent is beneficial in the optimization of neural networks (Goodfellow *et al.*, 2016); this is distinct from its use in the work of Chen & Oliver (2012); Emerick & Reynolds (2013) where it is used to facilitate posterior sampling, not optimization. Regarding the forward model we set $G_n^{(j)}(\cdot) = G_T(\cdot; x_0^{(j,n)})$. This approximate evaluation of $\mathcal{G}$, $G_n^{(j)}$, is found from choosing the initial condition $x_0^{(j,n)}$ at random and i.i.d with respect to both ensemble member j and iteration step n. Details are given for each example in what follows. Note that the finite time T in $G_n^{(j)}$ is typically different from τ arising in the time-averaged data.

Here we use the algorithm in settings where the number of parameters is small or similar to the dimension of the data, and regularization is not needed to prevent overfitting. The EKI algorithm as stated preserves the linear span of the initial ensemble $\{\theta_0^{(j)}\}_{j=1}^J$ for each n and thus operates in a finite dimensional vector space (Iglesias *et al.*, 2013, Theorem 2.1), even if Θ is an infinite dimensional vector space; this form of regularization may be useful for high-dimensional unknown parameters. Other types of regularization can also be incorporated into EKI if needed, as detailed in the introduction; in addition sparsity can be used to regularize as demonstrated in Schneider *et al.* (2020).

3.2 *Gaussian process regression*

Throughout this paper, we apply the EKI algorithm in a finite-dimensional vector space $\Theta \subseteq \mathbb{R}^p$. In some cases, we use GPR to estimate unknown functions appearing in our SDE. More precisely, we will use the mean $m(x)$ of a Gaussian process conditioned on noisy observations at a fixed set of design points; we will then optimize over the observed values of the process at the design points, over the standard deviations of the noise in these observations and over the parameters describing the covariance function of the Gaussian process. This leads to a parameter dependent function $m(x; \theta^{\mathrm{GP}})$,

the construction of which we now detail. Note that the construction involves probabilistic considerations but, once completed, leads to a class of deterministic functions $m(x;\theta^{\mathrm{GP}})$ over $\theta^{\mathrm{GP}} \in \Theta^{\mathrm{GP}}$. The key advantage of the GPR construction is that it leads to a hierarchical parameterization that has proved very useful in many machine learning tasks (see, e.g., Bernardo *et al.*, 1998; Rasmussen & Williams, 2006). Adapting it to statistical estimation more generally is potentially very fruitful and is one of the ideas we use here. We are essentially proposing to use Gaussian process parameterizations beyond the simple regression setting, into more general function learning problems.

The function m is defined as the minimizer of

$$L(m) := \frac{1}{2}\|m\|_{\mathsf{K}}^2 + \sum_{r=1}^{R} \frac{1}{2\sigma_{(r)}^2} |m(x_{(r)}) - m'_r|^2$$

over K, the reproducing kernel Hilbert space associated with covariance function $k(x, x'; a, \ell)$; here a and ℓ represent and amplitude and lengthscale parameter of the covariance function. By the representer theorem (see, e.g., Rasmussen & Williams, 2006), it follows that the minimizer lies in the linear span of $\{k(x, x_{(r)}; a, \ell)\}_{r=1}^{R}\}$ and may be found by solving a linear system of dimension equal to R. The solution of this linear system depends on the $\{m(x_{(r)}), \sigma_{(r)}\}_{r=1}^{R}$.

As a consequence, the set of parameters θ^{GP} defining our parameterized function $m(x;\theta^{\mathrm{GP}})$ contains the following elements:

(i) noisily observed values of $\{m(x_{(r)})\}$, at some fixed nodes $x_{(r)}$, comprising the vector $\{m'_r\}$;

(ii) observation error variances $\Sigma_{\mathrm{obs}} = \mathrm{diag}\{(\sigma_{(r)})^2\}$ at the fixed nodes $x_{(r)}$;

(iii) hyper-parameters (a, ℓ) that represents an amplitude and a length-scale of the kernel $k(x, x'; a, \ell)$ used in the GP regression.

In this setting m' and Σ_{obs} each contain R elements. Thus, $\Theta^{\mathrm{GP}} = \mathbb{R}^{2R+2}$. The result of the minimization is a complex nonlinear function of $\theta^{\mathrm{GP}} = (m', \Sigma_{\mathrm{obs}}, \sigma, \ell) \in \Theta^{\mathrm{GP}}$. We use EKI to learn θ^{GP} together with other unknown parameters in the modeled systems using time-averaged statistics as data. The linear span of $\{k(x, x_{(r)}; a, \ell)\}_{r=1}^{R}\}$ comprises a set of adaptive basis functions, which, via the parameters a and ℓ, may be adapted to the observed finite-time average data. The setting may be generalized or simplified in different ways: for example Σ_{obs} can be chosen to be an arbitrary symmetric positive-definite matrix to be learned or, as we do in the numerical examples in this paper, can be chosen as $\Sigma_{\mathrm{obs}} = \sigma^2\mathrm{I}$ where σ is optimized over and I denotes the identity matrix. The detailed settings of numerical experiments are described at the beginning of each sub-section in Section 4. Since we take Σ_{obs} to be a constant diagonal matrix throughout this paper, Θ^{GP} simplifies to $\mathbb{R}^{R+3}$.

4. Numerical results

To demonstrate the capability of the proposed methodology, we apply it to four different examples, including classical chaotic systems (Lorenz 63 system in subsection 4.1, Lorenz 96 system in subsection 4.2), climate dynamics (ENSO in subsection 4.3) and molecular dynamics (butane molecule dihedral angle in subsection 4.4). All these numerical studies confirm that the proposed methodology serves well as a systematic approach to the fitting of SDE models to data. In particular, the data we use appears to lead to identifiable parameter estimation problems in every example presented. In all cases, the data are initially presented in two figures, one showing the ability of the EKI method to fit the data,

and a second showing how well the fitted SDE performs in terms of reproducing the invariant measure of the true system. The ensemble size is chosen as 100 by default. The noise Γ is estimated based on the ensemble of time-averaged data from the system with random initial conditions. It should be noted that the time-averaged data (which serve as training data) are obtained from a relatively short trajectory, and the invariant measures (which serve as testing data) are obtained by simulating the modeled system for a much longer time. We then show other figures that differ from case to case and are designed to illustrate the quality and nature of the fitted SDE model.

4.1 *Lorenz 63 system*

4.1.1 *Noisy Lorenz 63: simulation study.* The first sets of experiments are simulation studies in which data from (2.4) are used to fit parameters within the following model:

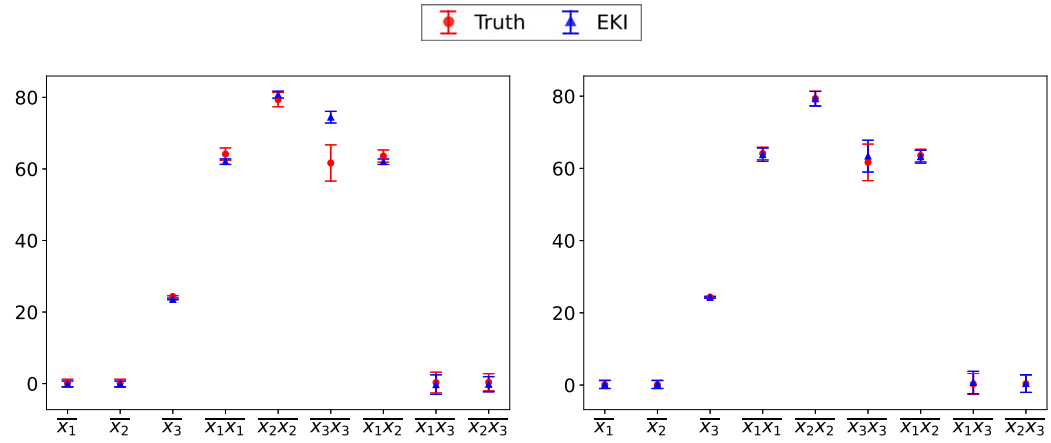
$$\begin{aligned} \frac{dx_1}{dt} &= \alpha(x_2 - x_1) + \sqrt{\sigma}\frac{dW_1}{dt} \\ \frac{dx_2}{dt} &= x_1(\rho - x_3) - g_L(x_2) + \sqrt{\sigma}\frac{dW_2}{dt} \\ \frac{dx_3}{dt} &= x_1x_2 - \beta x_3 + \sqrt{\sigma}\frac{dW_3}{dt}. \end{aligned} \tag{4.1}$$

Unlike the other examples in this paper, this initial simulation study involves data that come directly from an SDE within the model class being fitted, enabling a clear verification of the proposed methodology, and confirming that the time-averaged data lead to an identifiable model. Specifically, the data are obtained by simulating the model with a given set of parameters: $\alpha = 10$, $\rho = 28$, $\beta = 8/3$, and $\sigma = 10$, as well as the choice $g_L(x_2) = x_2$. Using EKI, we will fit θ defined in two different ways:

(i) Fix $g_L(x_2) = x_2$ and learn $\theta = (\alpha, \sigma)$. The initial ensemble of α and $\sqrt{\sigma}$ are uniformly drawn from $[1, 20]$ and $[0.1, 15]$.

(ii) Parameterize $g_L(x)$ by θ_1, as a GP, and learn $\theta = (\theta_1, \sigma)$. The initial ensemble of $\sqrt{\sigma}$ is uniformly drawn from $[0.1, 15]$. The fixed GP nodes are five points uniformly distributed in $[-30, 30]$. The initial ensemble of noisily observed values on those nodes are uniformly drawn from $[-20, 20]$. The initial ensemble of GP observation error is uniformly drawn from $[0.1, 10]$, and the initial ensemble of GP hyper-parameters a and ℓ are uniformly drawn from $[0.1, 10]$ and $[5, 10]$. Results are presented in Figs 8 to 10.

The trajectory initial condition is uniformly drawn from $[0, 1)$ for each state variable. 20 EKI iterations is used. Results are presented in Figs 4 to 5. The data in this case are a finite-time average approximation of $\{\mathcal{G}_1(x), \mathcal{G}_2(x)\}$, i.e., the first and second moments of the state vector x are used as observational data. Therefore, the data vector y has nine elements in total. In case (i), we fit both an ODE (constrain $\sigma = 0$ and learn α) and an SDE (learn both α and σ) to the given data. We are fitting one parameter in the ODE case and two parameters in the SDE setting, in both cases using a data vector y of dimension 9. The comparison of the output of the EKI algorithm with the true data, in the case of both ODE and SDE fits, is presented in Fig. 4. In the ODE case, the algorithm fails to match one second moment (left panel), while in the SDE case all moments are well-matched. This has a significant effect on the ability to reproduce the invariant measure (Fig. 5). In the upper row (ODE), the fit is very poor whereas in the lower row (SDE) it is excellent. The fit of the SDE is not surprising since the data are



(a) ODE (b) SDE

FIG. 4. First two moments of state x for the Lorenz 63 system found by using EKI to estimate α (ODE case) and σ, α (SDE case).

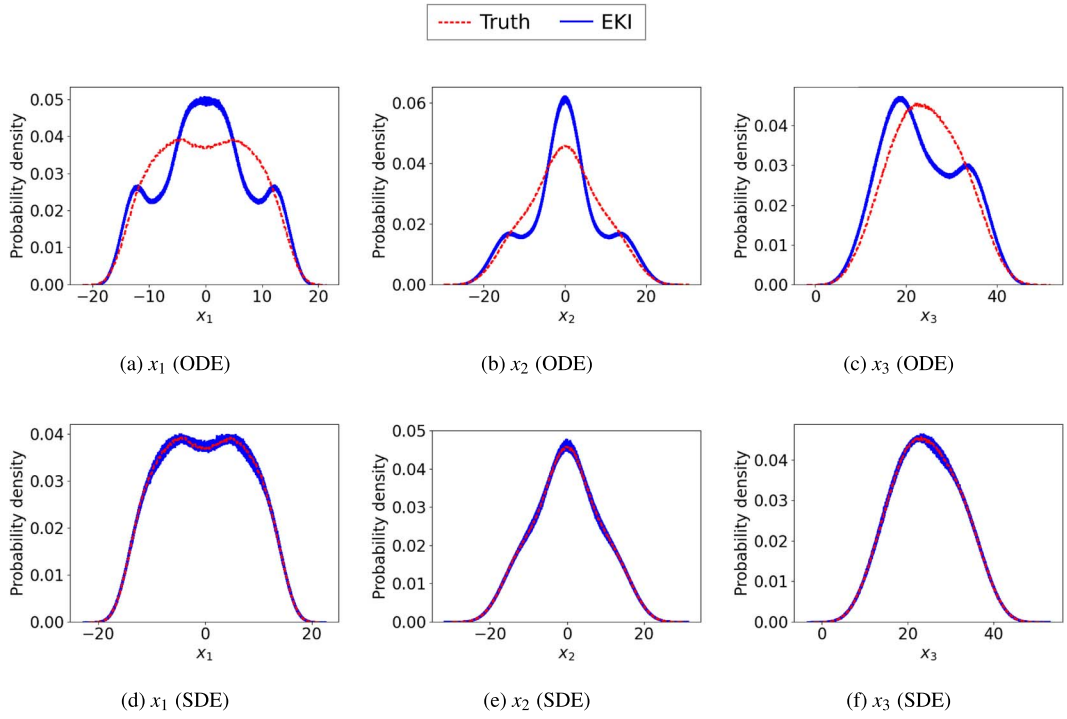
generated directly from within the model class to be fit; the behavior of the fit to an ODE gives insight into the method when the model is misspecified.

Figure 6 presents the trajectory of x_1 and the RMSE calculated using all state variables. To facilitate the comparison against the true system, we use the same initial condition, and the same random seed for the stochastic process, for both true and modeled systems. It should be noted that the training of the modeled systems does not make use of trajectory information from the true system. As shown in Fig. 6, the modeled system with the stochastic term demonstrates better performance in matching the trajectory of the true system.

The comparison of the PSD is presented in Fig. 7. Although the PSD is not used as data in this example, both fitted models capture the general pattern of the PSD of the true system. For more complex systems, PSD can provide additional information to help better identify the modeled system, and we demonstrate the use of PSD as part of data in Sections 4.3 and 4.4.

We now perform exactly the same set of experiments, fitting both an ODE and an SDE, but allowing the function g_L to be learnt as well. The function g_L is parameterized by the mean of a GP (with unknown values specified at five fixed nodes, and unknown constant observation error and two unknown hyper-parameters). Thus we are fitting eight parameters for the ODE and nine parameters for the SDE, using a data vector y of dimension 9. When we do this, we are able to obtain a good fit for the ODE in both data space, seen in Fig. 8a, and as measured by the fit to the invariant measure, as seen in Fig. 9(a–c). Nonetheless, the fit achieved by using an SDE remains substantially better, as seen in Fig. 8(b) and Fig. 9(d–f). The ensemble means of the Gaussian process learnt in the ODE and SDE models are presented in Fig. 10. Both GPs successfully capture the linear trend in the middle range of x_2, while the GP learnt in the ODE model exhibits more oscillations around the middle part and more rapid deviation from the linear trend at both ends.

4.1.2 *Deterministic Lorenz 63: dimension reduction.* We fit a two-dimensional SDE model of the form (2.7) to data generated from the first two components of the three-dimensional ODE model (2.5), (2.6). The four unknown functions $(\psi_1(\cdot), \psi_2(\cdot), \sigma_1(\cdot), \sigma_2(\cdot))$ are parameterized by the mean

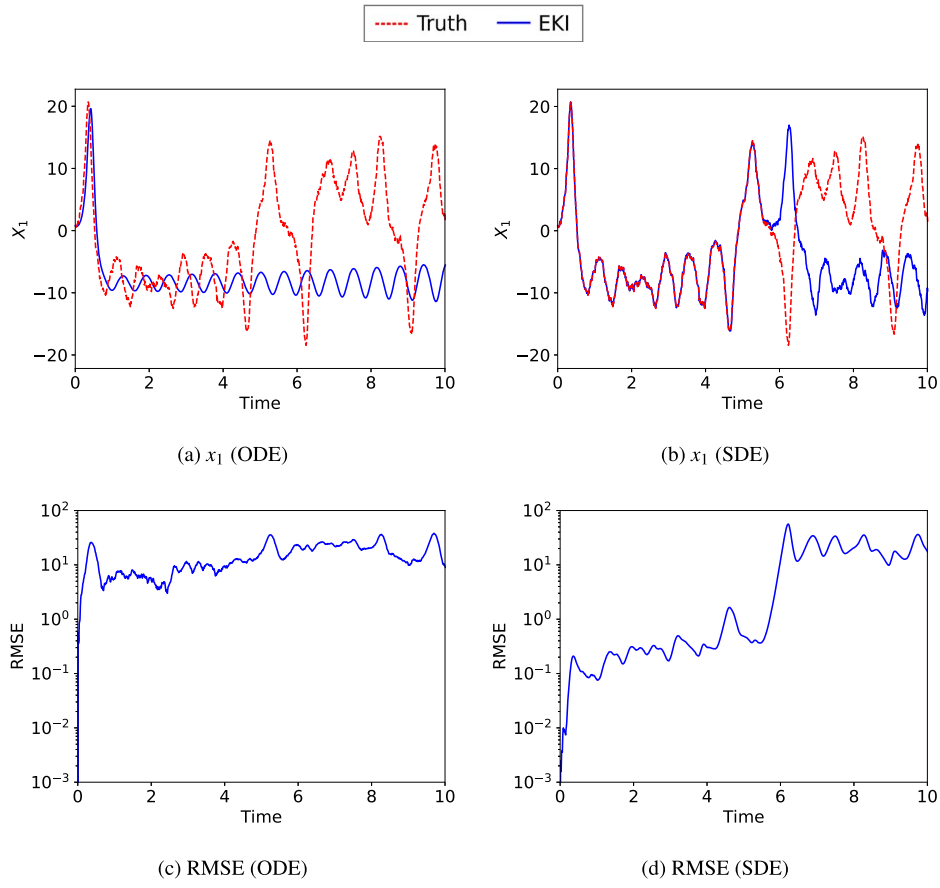


(a) x_1 (ODE) (b) x_2 (ODE) (c) x_3 (ODE)

(d) x_1 (SDE) (e) x_2 (SDE) (f) x_3 (SDE)

FIG. 5. Invariant measures of fitted models compared with those of the true noisy Lorenz 63 system: ODE case, fit α; SDE case, fit σ, α.

of GPR. Each GPR involves the mean values at five fixed nodes, a stationary observation error, and the stationary hyper-parameters $(\sigma_{\mathscr{G}\mathscr{P}}, \ell)$. The precise form of the data in this case is a finite time average approximation of $\{\mathscr{G}_m(a), \mathscr{G}_{ac}(a)\}$. Specifically, the first four moments of the state vector a are used in this case. In addition, nine equally spaced data points are used from a finite time average approximation of the autocorrelation function $\mathscr{G}_{ac}(a)$ for both states a_1 and a_2, from an interval of 10 time units, and we obtain 18 data points in total from the autocorrelation function. Thus we are fitting 32 parameters for the SDE by using a data vector y of dimension 27. The fixed GP nodes are five points uniformly distributed in $[-30, 30]$. The initial ensemble of noisily observed values on those nodes is uniformly drawn from $[-20, 20]$. The initial ensemble of GP observation error is uniformly drawn from $[0.1, 10]$, and the initial ensemble of GP hyper-parameters a and ℓ is uniformly drawn from $[0.1, 10]$ and $[5, 10]$. The trajectory initial condition is uniformly drawn from $[0, 1)$ for each state variable. We use 30 EKI iterations. The true moment data and autocorrelation data are presented in Fig. 11, alongside the results obtained by using EKI to fit the SDE model to this data, showing a relatively good match to the data. When measured by the ability to reproduce the invariant measure, the results demonstrate a strong match between the marginal invariant densities on a_1 and a_2 from the original three-dimensional ODE and from the fitted two-dimensional SDE (Fig.12). The fit to the autocorrelation functions of a_1 and a_2 is also quite good (Fig. 13). It should be noted that Fig. 11 presents the autocorrelation at discretized time shifts where we have training data, and Fig. 13 presents the autocorrelation which also extends beyond the time shifts used in training. Figure 14 compares the trajectories of the three-dimensional ODE (2.5),



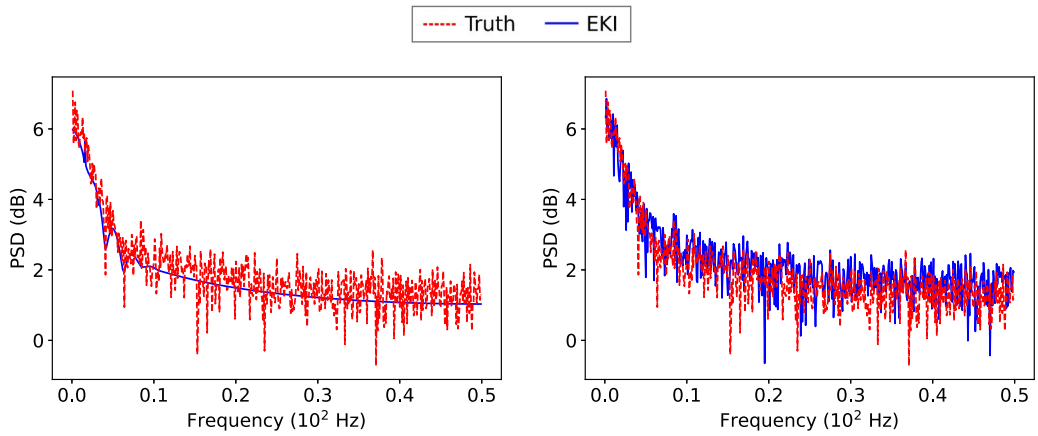
(a) x_1 (ODE) (b) x_1 (SDE)

(c) RMSE (ODE) (d) RMSE (SDE)

FIG. 6. State trajectory and RMSE of fitted models compared with those of the true noisy Lorenz 63 system: ODE case, fit α; SDE case, fit σ, α. The results are obtained by fitted models with ensemble mean of estimated parameters.

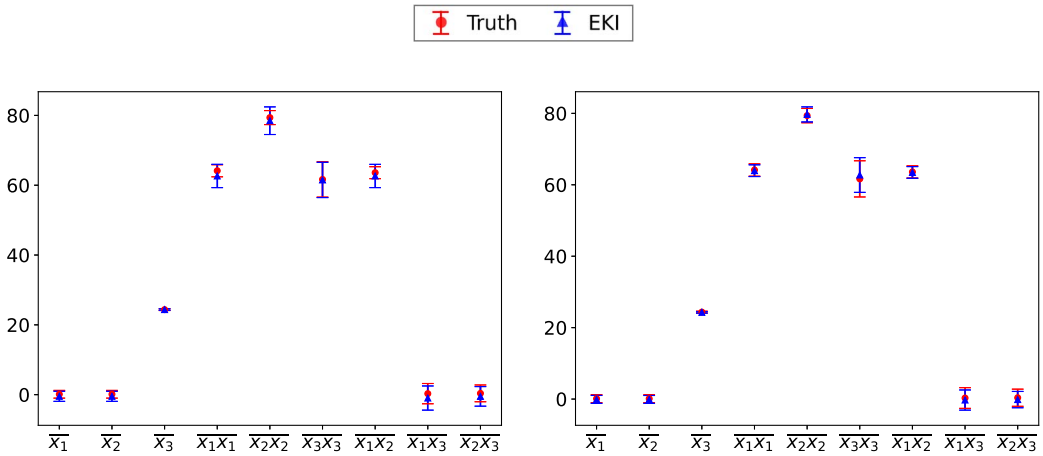
(2.6), projected on a_1 and a_2, with those of the fitted SDE (2.7); the difference in smoothness of the solutions of ODEs and SDEs is apparent at that level, although it can be seen that the fitted SDE model shows a similar pattern of irregular switching between the two distinct components of the attractor in the true system.

Despite the seemingly simple formulation of this example as outlined in section 2, it is a rather difficult problem. On the one hand, the transformed Lorenz 63 system (2.5), (2.6) is well known as a classical chaotic system. On the other hand, chaos cannot arise in the reduced-order two-dimensional model of the form (2.7) in a deterministic setting where the noise levels σ_i are set to zero, by the Poincaré–Bendixson theorem. Palmer (2001) showed that a reduced-order two-dimensional model of the Lorenz 63 system in discrete time could exhibit *qualitative* features of the original three-dimensional system, if subjected to additive noise, but it remained an unsolved problem to fit a stochastic two-dimensional model that *quantitatively* captures detailed chaotic behavior of the original three-dimensional chaotic system. This motivated us to introduce both more unknown parameters and more observed statistics as data, compared to the previous simulation study presented in (4.1.1). In so doing,



(a) ODE (b) SDE

FIG. 7. Power spectral density of fitted models compared with those of the true noisy Lorenz 63 system: ODE case, fit α; SDE case, fit σ, α. The results are obtained by fitted models with ensemble mean of estimated parameters.



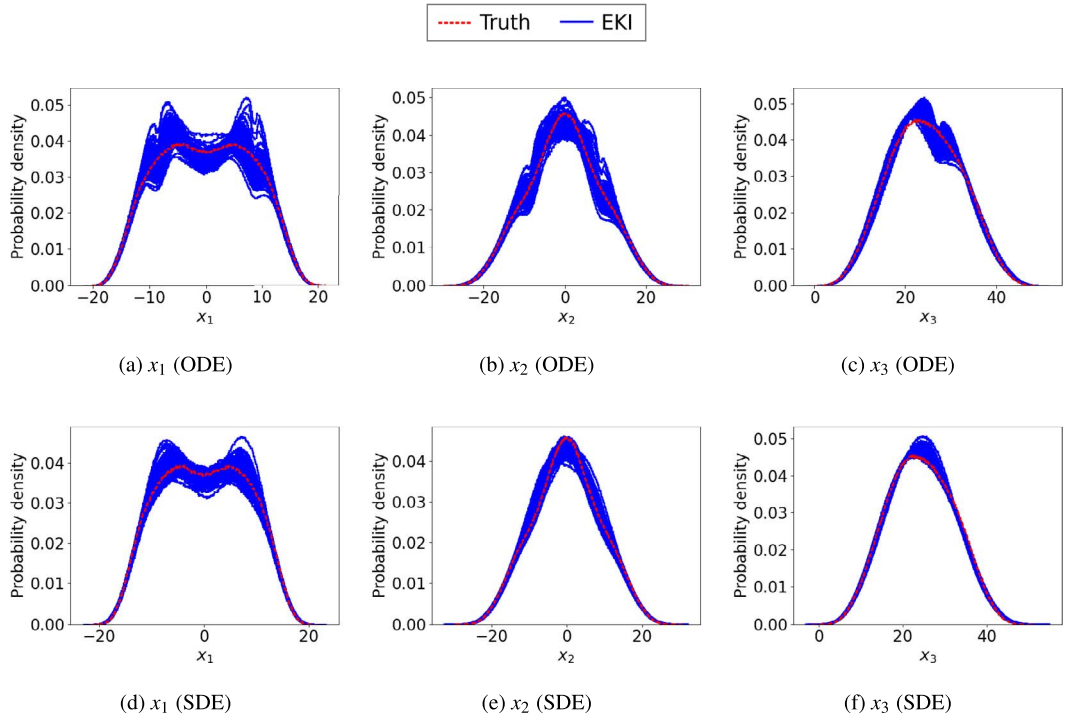
(a) ODE (b) SDE

FIG. 8. First two moments of state x for the Lorenz 63 system by using EKI to estimate σ and the linear function $g_L(x_2)$.

we have demonstrated some success in fitting an SDE model to data from the ODE. We face similar challenges in the remaining examples in this section, where the data are generated from a more complex model than the model that it is fitted, or indeed is actual observed data.

4.2 *Lorenz 96 system*

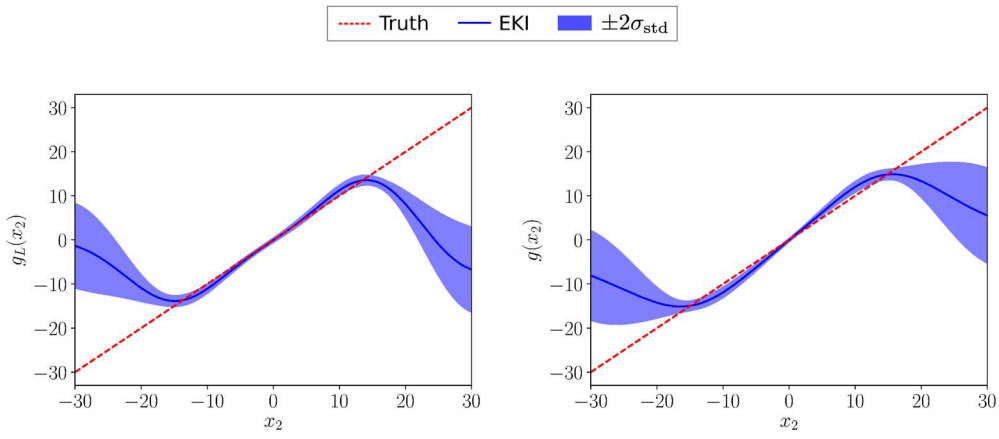
We generate data from the slow variable $\{x_k\}_{k=1}^K$ in (2.8) and use it to fit a parameterized function $\psi(\cdot)$ and parameter σ appearing in (2.11). In the experiments presented below, we take $K = 36$. The form



(a) x_1 (ODE) (b) x_2 (ODE) (c) x_3 (ODE)

(d) x_1 (SDE) (e) x_2 (SDE) (f) x_3 (SDE)

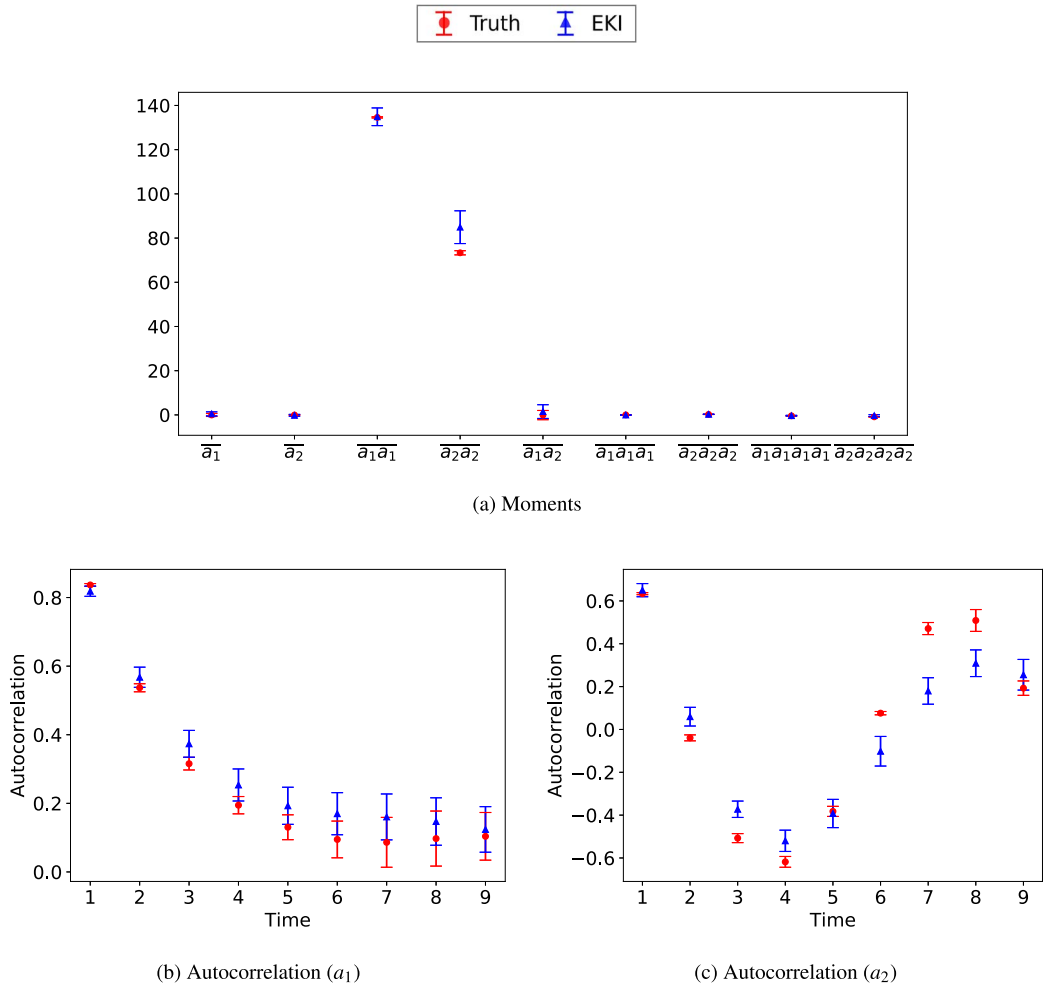
FIG. 9. Invariant measures of fitted models (with linear function $g_L(x_2)$ and σ as unknowns) and the true noisy Lorenz 63 system.



(a) ODE (b) SDE

FIG. 10. The linear function $g_L(x_2)$ learnt in (a) ODE model and (b) SDE model.

of the data in case (a) is finite-time averaged approximations of $\{\mathscr{G}_1(x), \mathscr{G}_2(x)\}$, i.e., observations of the first and second moments of the state vector $x = \{x_k\}_{k=1}^n$. The model discrepancy term $\psi(X_k)$ in (2.11)



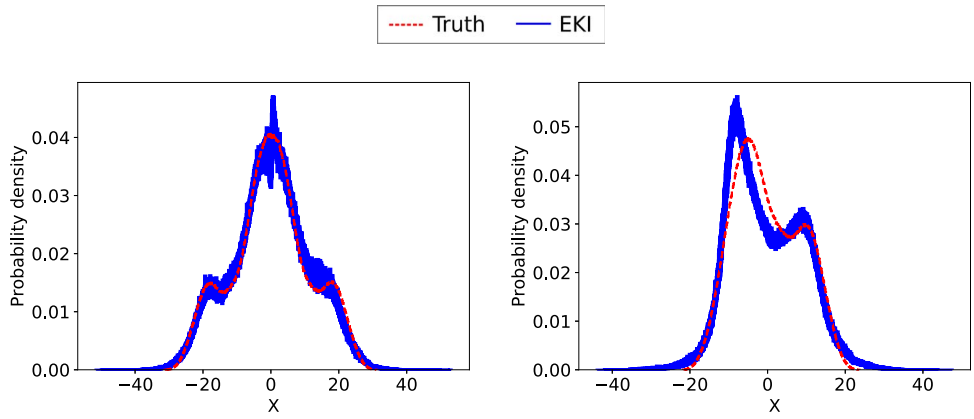
(a) Moments

(b) Autocorrelation (a_1)

(c) Autocorrelation (a_2)

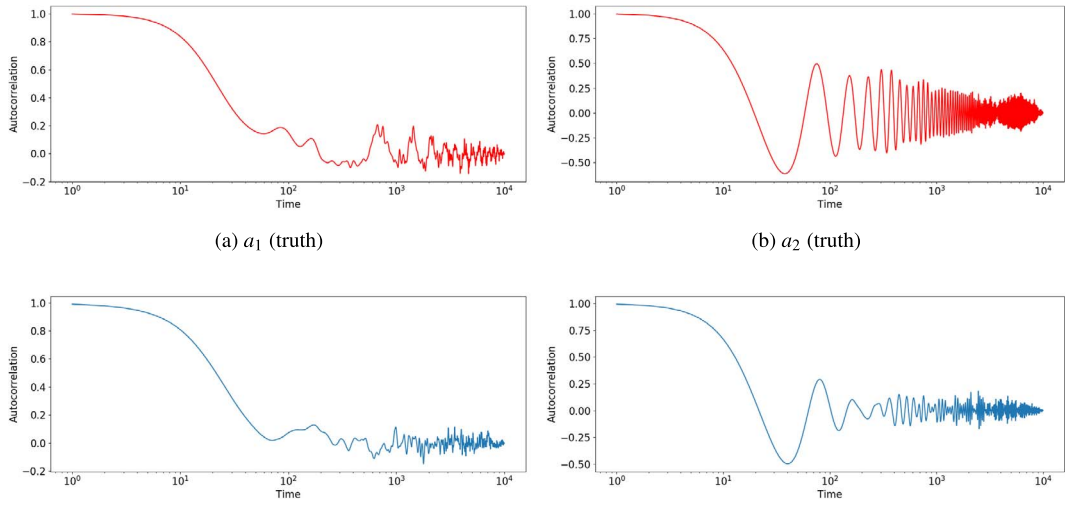
FIG. 11. Comparison of observation data between the true system and the EKI-fitted SDE model for the deterministic Lorenz 63 system.

is parameterized by the mean of a GP with mean values at 7 fixed nodes, and constant observation error and hyper-parameters. Thus, we are fitting 10 parameters for the ODE and 11 for the SDE, using a data vector y of dimension 44 (when only observing the first 8 slow variables). The form of the data in case (b) is finite-time averaged approximations of moments up to fourth order (only evaluating for each single slow variable and thus providing 144 data points) and 11 points on the averaged autocorrelation, leading to a data vector y of dimension 155. The model discrepancy term $\psi(X_k)$ in (2.11) is parameterized by the mean of a GP with mean values at 11 fixed nodes, and constant observation error and hyper-parameters. It should be noted that we also learn h^2c/J in case (b). Thus, we are fitting 15 parameters for the ODE and 16 for the SDE. The ensemble size is chosen as 300 for case (b). Our numerical results illustrate several interesting properties of stochastic closures for the Lorenz 96 multiscale system:



(a) a_1 (b) a_2

FIG. 12. Invariant measures of the deterministic Lorenz 63 system and the fitted reduced-order SDE model.



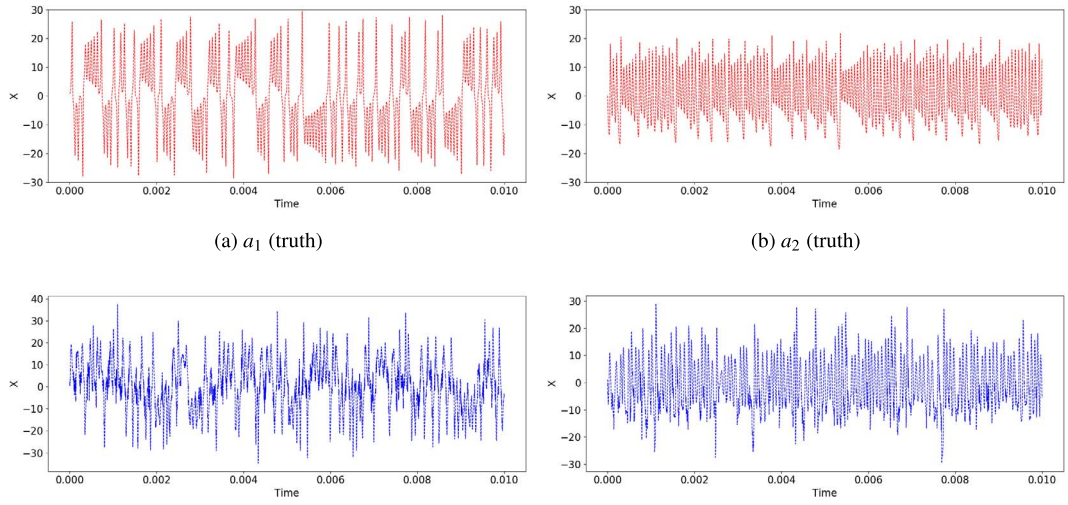
(a) a_1 (truth) (b) a_2 (truth)

(c) a_1 (EKI) (d) a_2 (EKI)

FIG. 13. Autocorrelation of the deterministic Lorenz 63 system and the fitted reduced-order SDE model.

(a) we show that for the relatively large time scale separation of $c = 10$ it is possible to fit both an accurate ODE ($\sigma = 0$) and SDE ($\sigma > 0$), in the sense that both ODE and SDE fitted models (2.8) accurately reproduce the invariant measure of the full system (2.8), using only $n = 8$;

(b) we show that with weaker scale separation of c $= 3$, the ODE fit is very poor, but for the SDE it becomes excellent. In this case, we still set $h = 1$ in the dynamics of fast variables for the full system.



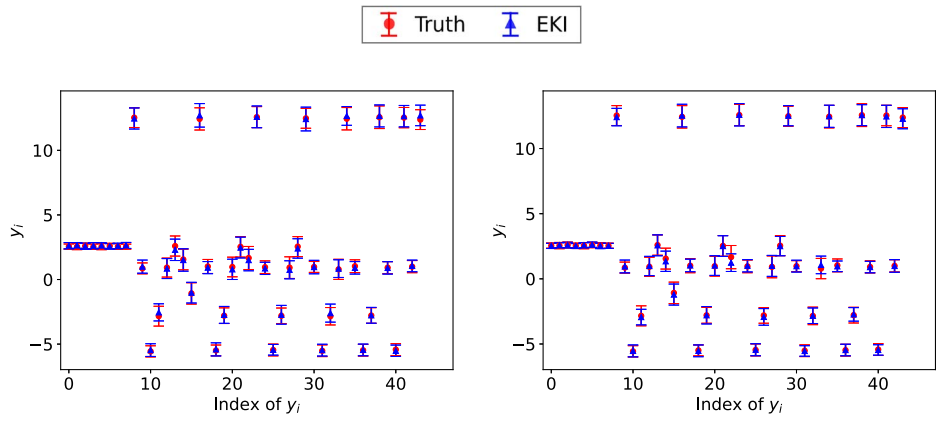
(a) a_1 (truth) (b) a_2 (truth)

(c) a_1 (EKI) (d) a_2 (EKI)

FIG. 14. Time series of the deterministic Lorenz 63 system and the fitted reduced-order SDE model.

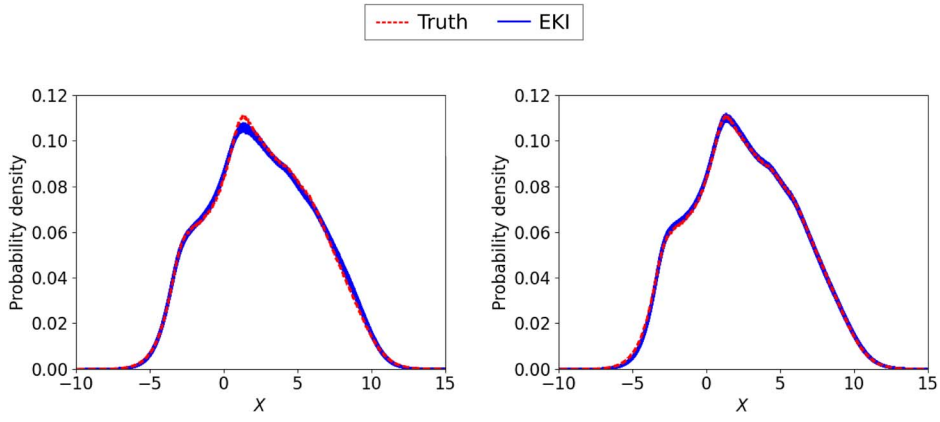
In case (a), the fixed GP nodes are seven points uniformly distributed in $[-15, 15]$. The initial ensemble of noisily observed values on those nodes is uniformly drawn from $[-1, 1]$. The initial ensemble of GP observation error is uniformly drawn from $[0.1, 1]$, and the initial ensemble of GP hyper-parameters a and ℓ is uniformly drawn from $[0.1, 1]$ and $[5, 20]$. In the SDE case, the initial ensemble of $\sqrt{\sigma}$ is drawn from $[0.01, 10]$. The trajectory initial condition is uniformly drawn from $[0, 1)$ for each state variable. 20 EKI iterations is used. Results for case (a) are presented in Figs 15 and 16. It can be seen in Fig. 15 that both the fitted ODE model and SDE model show almost perfect agreements with the true system in data space. Furthermore, both fitted ODE model and SDE model agree well with the true system when we compare the invariant measure with that of the underlying data-generating model, in Fig. 16. Although the agreement is slightly better for the fitted SDE model in Fig. 16b, the performances of fitted ODE and SDE models are quantitatively close to each other. The good performance of the fitted ODE model can be explained by invoking the averaging hypothesis, as proposed in Fatkullin & Vanden-Eijnden (2004), suggesting a closed ODE model of the form (2.11). The data implying the form of the closure ψ is as presented in Fig. 17a. Specifically, the scattering of the true closure term is narrower for $c = 10$, indicating that a deterministic closure of slow variables can achieve good agreement with the true system.

In case (b), the fixed GP nodes are 11 points uniformly distributed in $[-15, 15]$. The initial ensemble of noisily observed values on those nodes are uniformly drawn from $[-1, 1]$. The initial ensemble of GP observation error is uniformly drawn from $[0.1, 1]$, and the initial ensemble of GP hyper-parameters a and ℓ is uniformly drawn from $[0.1, 1]$ and $[1, 10]$. The initial ensemble of h^2c/J is uniformly drawn from $[1/3, 20/3]$. In the SDE case, the initial ensemble of $\sqrt{\sigma}$ is drawn from $[0.01, 10]$. The trajectory initial condition is uniformly drawn from $[0, 1)$ for each state variable. Twenty EKI iterations are used. Results for case (b) are presented in Figs 18 to 20. The averaging hypothesis does not apply so cleanly in this case where $c = 3$, as can be seen by comparing the moderate scattering in the data when $c = 10$ (Fig. 17a), with that obtained when $c = 3$ (Fig. 17b). In the second setting, it turns out that the fitted



(a) ODE (b) SDE

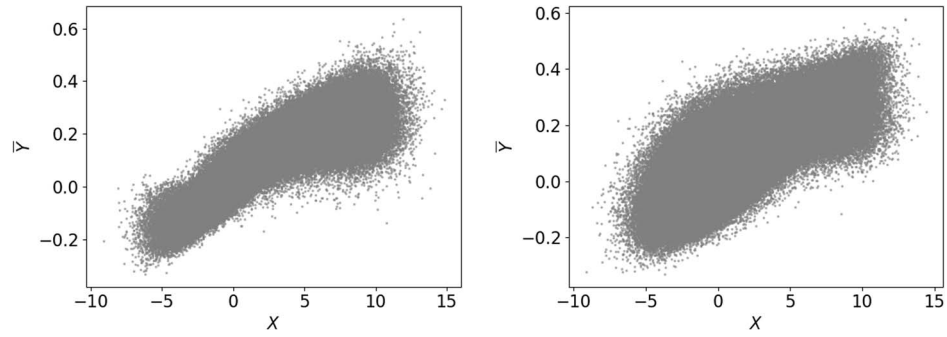
FIG. 15. Comparison of observation data between the true Lorenz 96 system ($c = 10$) and the fitted ODE and SDE models.



(a) ODE (b) SDE

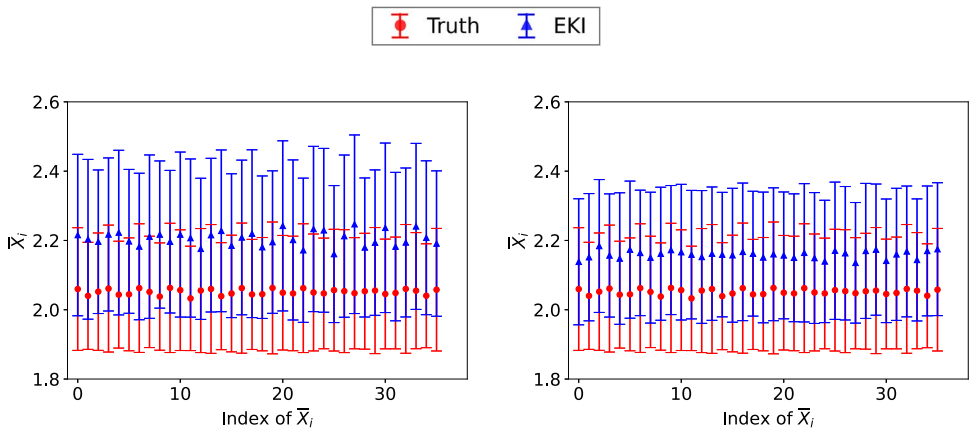
FIG. 16. Comparison of invariant measures between the true Lorenz 96 system ($c = 10$) and the fitted ODE and SDE models.

ODE model is far less satisfactory than the fitted SDE model. Figure 18 shows that the SDE fit achieves similar agreement as the ODE fit in data space. As presented in Fig. 19, the fitted ODE model still tends to concentrate toward a small number of discrete values in the long-time behavior, while the invariant measure of fitted SDE model demonstrates excellent agreement with the true system. The issue of the fitted ODE model in failing to maintain correct chaotic behavior for a long time can be more clearly seen in the time series presented in Fig. 20. More specifically, Fig. 20 demonstrates that the fitted ODE model is able to maintain the chaotic behavior within the time range $[0, 100]$ that is used to evaluate time-averaged statistics as data. Beyond that time range, the fitted ODE model is attracted to a quasi-periodic regime and then stays in that regime afterwards.



(a) $c = 10,\ h = 1$ (b) $c = 3,\ h = 10/3$

FIG. 17. True closure term of slow variables with different temporal scale separations between slow and fast variables.

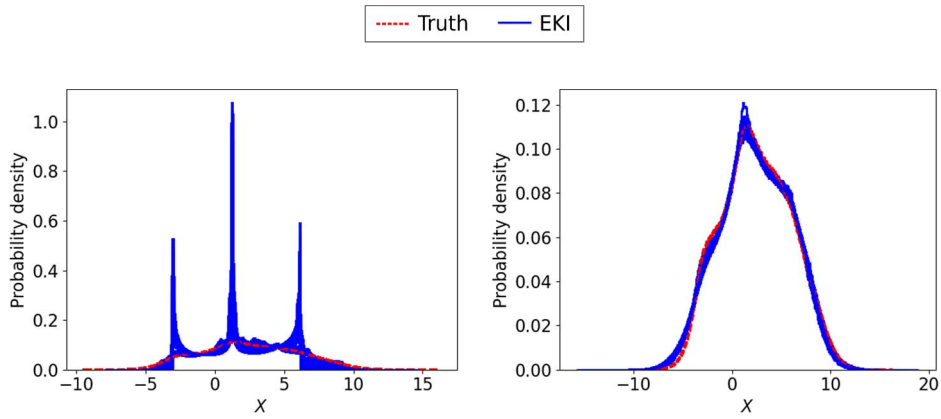


(a) ODE (b) SDE

FIG. 18. Comparison of data between the true Lorenz 96 system ($c = 3$) and the fitted ODE and SDE models. All 36 slow variables are used to compute the moments as data (only the first moment data are presented here).

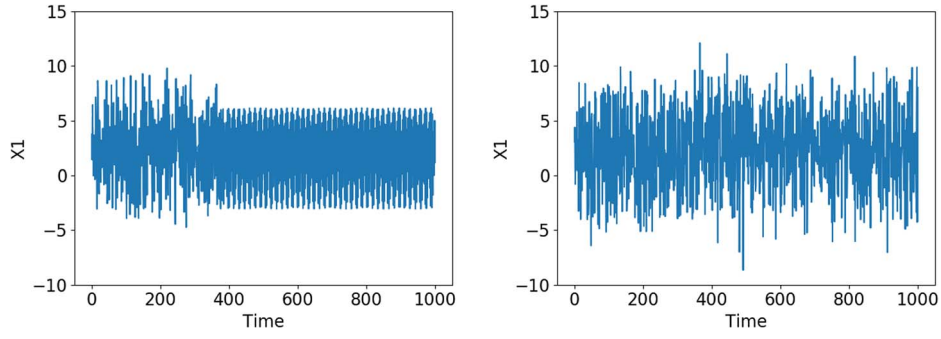
4.3 *El Niño–Southern oscillation*

In this subsection we use the time-series data (Rayner *et al.*, 2003) for SST T shown in Fig. 1(a) to fit an SDDE of the form (2.12), learning the four parameters a, b, c, σ. Recall that the delays τ_1 and τ_2 may be viewed as known properties. The precise form of the data in this case is a finite-time average approximation of $\{\{\mathcal{G}_m(T)\}_{m=1}^4, \mathcal{G}_{ac}(T), \mathcal{G}_{psd}(T)\}$. Thus, we are fitting four parameters in the SDDE by using a data vector y of dimension 16. The initial ensemble of a, b, c are drawn uniformly from $[0.1, 10]$. The initial ensemble of $\log(\sqrt{\sigma})$ is drawn uniformly from $[\log(0.1), \log(10)]$. The trajectory initial condition is uniformly drawn from $[0, 1)$ for each state variable. 10 EKI iterations is used. The true moment data are presented in Fig. 21(a). To use the autocorrelation function $\mathcal{G}_{ac}(T)$ as data, we sample nine points from it with an interval of six months as presented in Fig. 21(b). We also use the coefficients



(a) ODE (b) SDE

FIG. 19. Comparison of invariant measures between the true Lorenz 96 system ($c = 3$) and the fitted ODE and SDE models.



(a) ODE (b) SDE

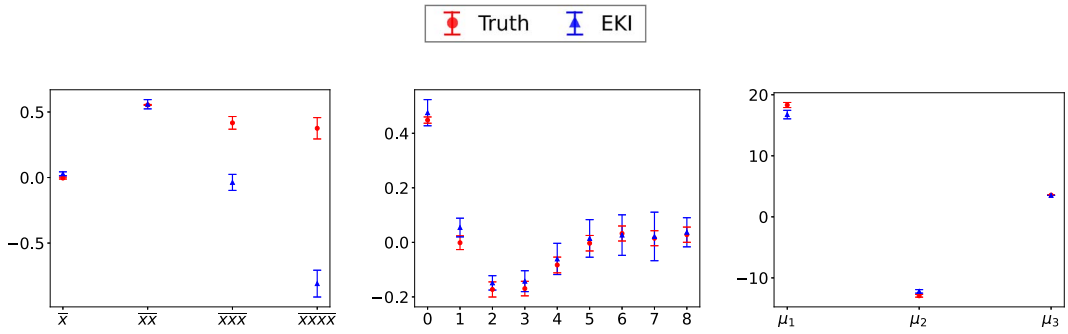
FIG. 20. Comparisons of time series of slow variable X_1 by using (a) the fitted ODE model and (b) the fitted SDE model. The time range over which we collect data for EKI is [0, 100].

of a second-order polynomial fit to the logarithm of the PSD; the three coefficients are presented in Fig. 21(c). Results demonstrating the fit are presented in Figs 21 to 22.

It can be seen in Fig. 21 that the fitted SDE model shows a good agreement with true ENSO statistics at first and second order, but not a higher order. Looking at the invariant measures presented in Fig. 22, we see evidence that the fitted SDE model can capture the long-time behavior of ENSO. The fitted SDE model does not provide a good agreement with data especially in the fourth-order moment, indicating the limitation of the current SDE model, which could be addressed by introducing a more sophisticated model.

4.4 *Butane molecule dihedral angle*

We fit the second-order Langevin equation (2.14) model to data derived from (2.13). The precise form of the data in this case is a finite-time average approximation of $\{\{\mathcal{G}_m(\phi)\}_{m=1}^4, \mathcal{G}_{ac}(\phi), \mathcal{G}_{psd}(\phi)\}$, where ϕ



(a) Moments (b) Autocorrelation (c) PSD

FIG. 21. The comparison of observed quantities, including (a) the first four moments, (b) the autocorrelation function and (c) the coefficients of fitted second-order polynomial of power spectrum density.

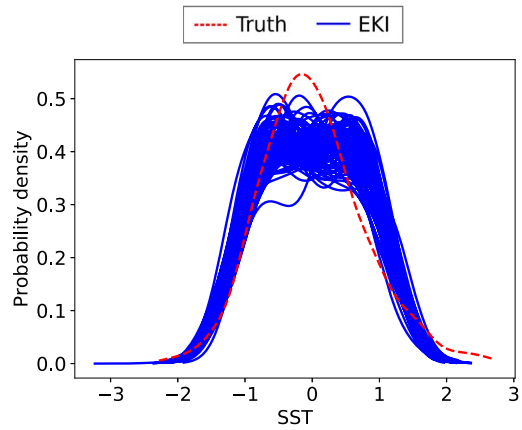
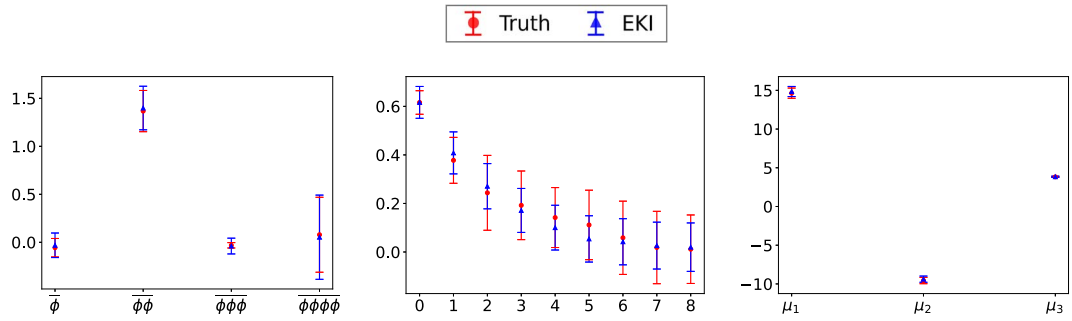


FIG. 22. The comparison of the invariant measures of SST between true ENSO data and fitted SDE model.

denotes the dihedral angle. Thus, similarly to the previous subsection, we use the first four moments of ϕ, and the true moment data are presented in Fig. 23(a). Furthermore, $\mathcal{G}_{ac}(\phi)$ is approximated using nine points sampled from the autocorrelation function with an interval of 50 fs ('fs' stands for femtosencond, and 1 fs $= 10^{-15}$ s), and these sampled data points are presented in Fig. 23(b). Finally, we also use the coefficients of a second-order polynomial that fits to the logarithm of the PSD, and the three coefficients are presented in Fig. 23(c). On the other hand, we are learning scalars γ and σ in (2.13), together with the potential Ψ constructed from Gaussian basis functions (with length scale fixed as 0.5) centered at nine points evenly distributed in $[-\pi, \pi]$. Thus, we are fitting 11 parameters for the SDE by using a data vector y of dimension 16. The initial ensemble of γ and $\log(\sqrt{2\sigma\gamma})$ are drawn uniformly from $[0.1, 2]$ and $[\log(0.1), \log(3)]$. The trajectory initial condition is uniformly drawn from $[0, 1)$ for each state variable. 10 EKI iterations is used. Results showing the fitted SDE are presented in Figs 23 to 26.

Figure 23 shows that the fitted SDE model achieves very good agreement with all the statistics of true dihedral angle data. More importantly, we can see in Fig. 24 that the fitted SDE model also leads to an invariant measure that agrees well with that of the true dihedral angle data.



(a) Moments (b) Autocorrelation (c) PSD

FIG. 23. The comparison of observed quantities of dihedral angle ϕ, including (a) the first four moments, (b) autocorrelation function and (c) coefficients of fitted second order polynomial of power spectral density.

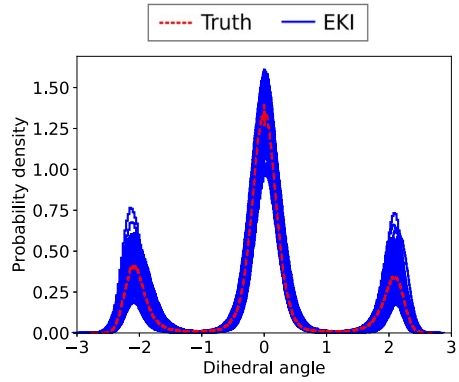
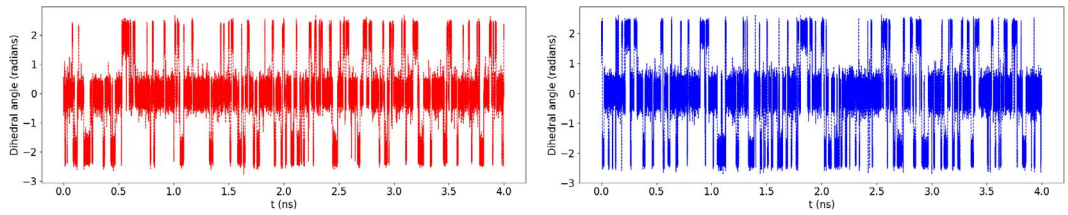


FIG. 24. Invariant measures of butane molecule dihedral angle.

In order to further validate the fitted SDE model, and in particular to show that it successfully captures the transition behavior and frequency information of the true data, we further study the time series (Fig. 25) and autocorrelation function (Fig. 26) by simulating the fitted SDE model for a long time. Figs. 25 and 26 clearly show that the fitted SDE model successfully reproduces the statistical behavior of the true dihedral angle computed from the much more expensive full molecular dynamics simulation.

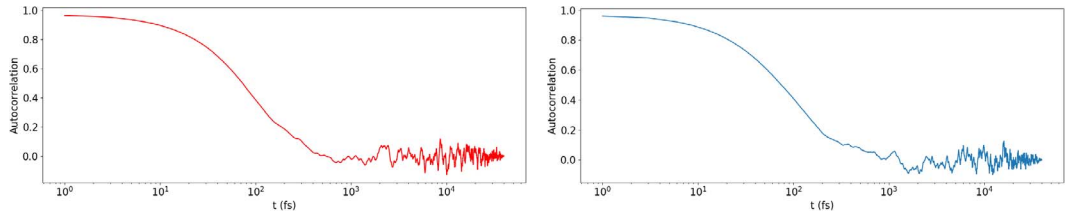
5. Conclusions

Although computing power has increased dramatically in the past several decades, so too has the complexity of models that scientists wish to simulate. It is still infeasible to resolve all the interactions within the true system in many applications. SDEs arise naturally as models in many disciplines, even if the first principles governing equations are deterministic, because stochasticity can effectively model unresolved variables. However, standard statistical methodology for parameter estimation is not always suitable for fitting SDE models to real data, due to the frequently observed inconsistency between SDEs and real data at small time scales. In this work, we exploit the idea of using sufficient statistics found by



(a) Truth (b) EKI

FIG. 25. Time series of butane molecule dihedral angle.



(a) Truth (b) EKI

FIG. 26. Autocorrelation of butane molecule dihedral angle.

finite-time averaging time series data. Using these statistics, we demonstrated that an SDE model, with the unknown functions being parameterized by GPR, can be fitted to the data using EKI methods built as optimization-based variants of the Bayesian sampling algorithms proposed in Chen & Oliver (2012) and Emerick & Reynolds (2013).

Ensemble methods are particularly well-suited to this problem for several reasons: they are derivative-free, thereby sidestepping computational and technical issues that arise from differentiating SDEs with respect to parameters; they are inherently parallelizable; they are robust to the use of approximate, noisy forward model evaluations; and they scale well to high-dimensional problems. Although differentiating SDEs with respect to parameters does not cause issues in special settings, e.g., when the diffusion coefficient is constant so that the stochastic process is not state-dependent, differentiation is more problematic when the diffusion coefficient is state-dependent. Unlike derivative-based optimization methods, ensemble methods avoid differentiating SDEs with respect to parameters and thus are applicable to more general settings. High-dimensional parameter learning would require regularization in combination with the basic EKI algorithm used here, as discussed in the introduction. The novel hierarchical GPR-based function approximation that we use meshes well with the EKI methodology.

Future directions of interest in this area include the derivation of a systematic approach to the determination of sufficient statistics, analysis of the EKI algorithm for learning in the context of these problems, and analysis of the use of GPR-based function representation in nonlinear inverse problems and the further use of the methodology to new problems arising in applications.

Acknowledgements

The authors thank Yvo Pokern (University College London) for providing the butane dihedral angle data and giving advice on using it. They are also grateful to Sebastian Reich for discussion of several aspects of the contents of this paper, leading to an improved presentation.

Funding

All authors are supported by the generosity of Eric and Wendy Schmidt by recommendation of the Schmidt Futures program; Earthrise Alliance; Mountain Philanthropies; the Paul G. Allen Family Foundation; the National Science Foundation (award AGS1835860, award DMS-1818977 to A.M.S.); Office of Naval Research (award N00014-17-1-2079).

References

Abdulle, A., Garegnani, G., Pavliotis, G. A., Stuart, A. M. & Zanoni, A. (2021) Drift estimation of multiscale diffusions based on filtered data. *Found. Comput. Math.*, 1–52.

Albers, D. J., Levine, M., Gluckman, B., Ginsberg, H., Hripcsak, G. & Mamykina, L. (2017) Personalized glucose forecasting for type 2 diabetes using data assimilation. *PLoS Comput. Biol.*, **13**.

Albers, D. J., Blancquart, P.-A., Levine, M. E., Esmaeilzadeh Seylabi, E. & Stuart, A. (2019) Ensemble Kalman methods with constraints. *Inverse Probl.*, **35**, 095007.

Anderson, J. L. (2001) An ensemble adjustment Kalman filter for data assimilation. *Mon. Weather Rev.*, **129**, 2884–2903.

Arnold, H., Moroz, I. & Palmer, T. (2013) Stochastic parametrizations and model uncertainty in the Lorenz'96 system. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, **371**, 20110479.

Bernardo, J., Berger, J., Dawid, A., Smith, A., et al. (1998) Regression and classification using Gaussian process priors. *Bayesian Stat.*, **6**, 475.

Bocquet, M., Brajard, J., Carrassi, A. & Bertino, L. (2020) Bayesian inference of chaotic dynamics by merging data assimilation, machine learning and expectation-maximization. *Found. Data Sci.*, **2**, 55–80.

Bocquet, M. & Sakov, P. (2012) Combining inflation-free and iterative ensemble Kalman filters for strongly nonlinear systems. *Nonlinear Process. Geophys.*, **19**, 383–399.

Bocquet, M. & Sakov, P. (2014) An iterative ensemble Kalman smoother. *Q. J. R. Meteorol. Soc.*, **140**, 1521–1535.

Boninsegna, L., Nüske, F. & Clementi, C. (2018) Sparse learning of stochastic dynamical equations. *J. Chem. Phys.*, **148**, 241723.

Brockwell, P. J., Davis, R. A. & Fienberg, S. E. (1991) *Time Series: Theory and Methods*. New York: Springer Science & Business Media.

Brunton, S. L. & Kutz, J. N. (2019) *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge, United Kingdom: Cambridge University Press.

Buckwar, E. (2000) Introduction to the numerical analysis of stochastic delay differential equations. *J. Comput. Appl. Math.*, **125**, 297–307.

Callaham, J. L., Loiseau, J.-C., Rigas, G. & Brunton, S. L. (2021) Nonlinear stochastic modelling with Langevin regression. *Proc. R. Soc. A*, **477**, 20210092.

Carrillo, J. A., Choi, Y.-P., Totzeck, C. & Tse, O. (2018) An analytical framework for consensus-based global optimization method. *Math. Models Methods Appl. Sci.*, **28**, 1037–1066.

Chada, N. K., Stuart, A. M. & Tong, X. T. (2020) Tikhonov regularization within ensemble Kalman inversion. *SIAM J. Numer. Anal.*, **58**, 1263–1294.

Chen, Y. & Oliver, D. (2012) Ensemble randomized maximum likelihood method as an iterative ensemble smoother. *Math. Geosci.*, **44**, 1–26.

Cleary, E., Garbuno-Inigo, A., Lan, S., Schneider, T. & Stuart, A. M. (2021) Calibrate, emulate, sample. *J. Comput. Phys.*, **424**, 109716.

Coifman, R. R., Kevrekidis, I. G., Lafon, S., Maggioni, M. & Nadler, B. (2008) Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems. *Multiscale Model. Simul.*, **7**, 842–864.

Cotter, C., Pavliotis, G., et al. (2009) Estimating eddy diffusivities from noisy Lagrangian observations. *Commun. Math. Sci.*, **7**, 805–838.

Diekmann, A. & Mitter, P. (2014) *Stochastic Modelling of Social Processes*. Cambridge, MA: Academic Press.

Djurdjevac, N., Sarich, M. & Schütte, C. (2010) *On Markov state models for metastable processes. Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*. World Scientific, pp. 3105–3131.

Doherty, J. E., Fienen, M. N. & Hunt, R. J. (2010) Approaches to highly parameterized inversion: pilot-point theory, guidelines, and research directions. *US Geologic. Survey Sci. Investig. Rep.*, **5168**, 36.

Doucet, A., De Freitas, N. & Gordon, N. (2001) An introduction to sequential Monte Carlo methods. *Sequential Monte Carlo Methods in Practice*. New York: Springer, pp. 3–14.

Dunbar, O. R., Garbuno-Inigo, A., Schneider, T. & Stuart, A. M. (2021) Calibration and uncertainty quantification of convective parameters in an idealized GCM. *J. Adv. Model. Earth Syst.*, **13**, e2020MS002454.

Duncan, A., Stuart, A. & Wolfram, M.-T. (2021) Ensemble inference methods for models with noisy and expensive likelihoods. arXiv preprint arXiv:2104.03384.

Emerick, A. A. & Reynolds, A. C. (2013) Ensemble smoother with multiple data assimilation. *Comput. Geosci.*, **55**, 3–15.

Engl, H. W., Hanke, M. & Neubauer, A. (1996) *Regularization of Inverse Problems*, vol. 375. Dordrecht, Netherlands: Springer Netherlands.

Erneux, T. (2009) *Applied Delay Differential Equations*, vol. **3**. New York: Springer Science & Business Media.

Evensen, G. (1994) Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res. Oceans*, **99**, 10143–10162.

Evensen, G. (2018) Analysis of iterative ensemble smoothers for solving inverse problems. *Comput. Geosci.*, **22**, 885–908.

Evensen, G. (2019) Accounting for model errors in iterative ensemble smoothers. *Comput. Geosci.*, **23**, 761–775.

Fatkullin, I. & Vanden-Eijnden, E. (2004) A computational strategy for multiscale systems with applications to Lorenz 96 model. *J. Comput. Phys.*, **200**, 605–638.

Fearnhead, P. & Prangle, D. (2012) Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *J. Royal Stat. Soc. Ser. B (Stat. Methodol.)*, **74**, 419–474.

Ferguson, A. L., Panagiotopoulos, A. Z., Kevrekidis, I. G. & Debenedetti, P. G. (2011) Nonlinear dimensionality reduction in molecular simulation: the diffusion map approach. *Chem. Phys. Lett.*, **509**, 1–11.

Frankignoul, C. & Hasselmann, K. (1977) Stochastic climate models, part ii application to sea-surface temperature anomalies and thermocline variability. *Tellus*, **29**, 289–305.

Froyland, G., Gottwald, G. A. & Hammerlindl, A. (2014) A computational method to extract macroscopic variables and their dynamics in multiscale systems. *SIAM J. Appl. Dynam. Syst.*, **13**, 1816–1846.

Garbuno-Inigo, A., Nüsken, N. & Reich, S. (2020a) Affine invariant interacting Langevin dynamics for Bayesian inference. *SIAM J. Appl. Dynam. Syst.*, **19**, 1633–1658.

Garbuno-Inigo, A., Hoffmann, F., Li, W. & Stuart, A. M. (2020b) Interacting Langevin diffusions: gradient structure and ensemble Kalman sampler. *SIAM J. Appl. Dynam. Syst.*, **19**, 412–441.

Gardiner, C. (2009) *Stochastic Methods*, vol. 4. Berlin: Springer.

Giannakis, D. (2019) Data-driven spectral decomposition and forecasting of ergodic dynamical systems. *Appl. Comput. Harmon. Anal.*, **47**, 338–396.

Goel, N. S. & Richter-Dyn, N. (2016) *Stochastic Models in Biology*. Cambridge, MA: Elsevier.

Goodfellow, I., Bengio, Y., Courville, A. & Bengio, Y. (2016) *Deep Learning*, vol. **1**. Cambridge: MIT Press.

Goodman, J. & Weare, J. (2010) Ensemble samplers with affine invariance. *Commun. Appl. Math. Comput. Sci.*, **5**, 65–80.

Gu, Y. & Oliver, D. S. (2007) An iterative ensemble Kalman filter for multiphase fluid flow data assimilation. *SPE J.*, **12**, 438–446.

Hasselmann, K. (1976) Stochastic climate models part i. Theory. *Tellus*, **28**, 473–485.

Hasselmann, K. (1988) Pips and pops: the reduction of complex dynamical systems using principal interaction and oscillation patterns. *J. Geophys. Res. Atmos.*, **93**, 11015–11021.

Iglesias, M. A., Law, K. J. & Stuart, A. M. (2013) Ensemble Kalman methods for inverse problems. *Inverse Probl.*, **29**, 045001.

Iglesias, M. A. (2015) Iterative regularization for ensemble data assimilation in reservoir models. *Comput. Geosci.*, **19**, 177–212.

Iglesias, M. A. (2016) A regularizing iterative ensemble Kalman method for PDE-constrained inverse problems. *Inverse Probl.*, **32**, 025002.

Julier, S., Uhlmann, J. & Durrant-Whyte, H. F. (2000) A new method for the nonlinear transformation of means and covariances in filters and estimators. *IEEE Trans. Automat. Control*, **45**, 477–482.

Kalliadasis, S., Krumscheid, S. & Pavliotis, G. A. (2015) A new framework for extracting coarse-grained models from time series with multiscale structure. *J. Comput. Phys.*, **296**, 314–328.

Klus, S., Nüske, F., Koltai, P., Wu, H., Kevrekidis, I., Schütte, C. & Noé, F. (2018) Data-driven model reduction and transfer operator approximation. *J. Nonlinear Sci.*, **28**, 985–1010.

Krumscheid, S., Pavliotis, G. A. & Kalliadasis, S. (2013) Semiparametric drift and diffusion estimation for multiscale diffusions. *Multiscale Model. Simul.*, **11**, 442–473.

Krumscheid, S., Pradas, M., Pavliotis, G. & Kalliadasis, S. (2015) Data-driven coarse graining in action: modeling and prediction of complex systems. *Phys. Rev. E*, **92**, 042139.

Kutoyants, Y. A. (2013) *Statistical Inference for Ergodic Diffusion Processes*. New York: Springer Science & Business Media.

Kwasniok, F. & Lohmann, G. (2009) Deriving dynamical models from paleoclimatic records: application to glacial millennial-scale climate variability. *Phys. Rev. E*, **80**, 066104.

Leimkuhler, B. & Reich, S. (2004) *Simulating Hamiltonian Dynamics*, vol. 14. Cambridge, United Kingdom: Cambridge University Press.

Li, G. & Reynolds, A. C. (2009) Iterative Ensemble Kalman Filters for Data Assimilation. *SPE J.*, 496–505.

Lorenz, E. N. (1963) Deterministic nonperiodic flow. *J. Atmospheric Sci.*, **20**, 130–141.

Lorenz, E. N. (1996) Predictability: a problem partly solved. *Proc. Seminar on Predictability*, vol. 1.

Lütkepohl, H. (2013) *Introduction to Multiple Time Series Analysis*. New York: Springer Science & Business Media.

Majda, A. J. & Harlim, J. (2012) *Filtering Complex Turbulent Systems*. Cambridge, United Kingdom: Cambridge University Press.

Majda, A. J. & Kramer, P. R. (1999) Simplified models for turbulent diffusion: theory, numerical modelling, and physical phenomena. *Phys. Rep.*, **314**, 237–574.

Maybeck, P. S. (1982) *Stochastic Models, Estimation, and Control*. Cambridge, MA: Academic Press.

Neumaier, A. & Schneider, T. (2001) Estimation of parameters and eigenmodes of multivariate autoregressive models. *ACM Trans. Math. Softw.*, **27**, 27–57.

Nott, D. J., Marshall, L. & Ngoc, T. M. (2012) The ensemble Kalman filter is an abc algorithm. *Stat. Comput.*, **22**, 1273–1276.

Palmer, T. N. (2001) A nonlinear dynamical perspective on model error: a proposal for non-local stochastic-dynamic parametrization in weather and climate prediction models. *Q. J. R. Meteorol. Soc.*, **127**, 279–304.

Papaspiliopoulos, O., Pokern, Y., Roberts, G. O. & Stuart, A. M. (2012) Nonparametric estimation of diffusions: a differential equations approach. *Biometrika*, **99**, 511–531.

Papavasiliou, A., Pavliotis, G. & Stuart, A. (2009) Maximum likelihood drift estimation for multiscale diffusions. *Stoch. Process. Appl.*, **119**, 3173–3210.

Pavliotis, G. A., Pokern, Y. & Stuart, A. M. (2012) Parameter estimation for multiscale diffusions: an overview. *Stat. Methods Stochast. Differ. Equ.*, **124**, 429.

Pavliotis, G. & Stuart, A. (2007) Parameter estimation for multiscale diffusions. *J. Stat. Phys.*, **127**, 741–781.

Penland, C. & Magorian, T. (1993) Prediction of Niño 3 sea surface temperatures using linear inverse modeling. *J. Climate*, **6**, 1067–1076.

Pokern, Y., Stuart, A. M. & Wiberg, P. (2009) Parameter estimation for partially observed hypoelliptic diffusions. *J. Royal Stat. Soc. Ser. B (Stat. Methodol.)*, **71**, 49–73.

Pulido, M., Tandeo, P., Bocquet, M., Carrassi, A. & Lucini, M. (2018) Stochastic parameterization identification using ensemble Kalman filtering combined with maximum likelihood methods. *Tellus A*, **70**, 1442099.

Rasmussen, C. E. & Williams, C. (2006) *Gaussian Processes for Machine Learning, vol. 1. MIT Press*, **39**, 40–43.

Rayner, N., Parker, D. E., Horton, E., Folland, C. K., Alexander, L. V., Rowell, D., Kent, E. & Kaplan, A. (2003) Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res. Atmos.*, **108**.

Reich, S. (2011) A dynamical systems framework for intermittent data assimilation. *BIT Numer. Math.*, **51**, 235–249.

Sakov, P., Oliver, D. S. & Bertino, L. (2012) An iterative EnKF for strongly nonlinear systems. *Mon. Weather Rev.*, **140**, 1988–2004.

Schillings, C. & Stuart, A. M. (2017) Analysis of the ensemble Kalman filter for inverse problems. *SIAM J. Numer. Anal.*, **55**, 1264–1290.

Schlick, T. (2010) *Molecular Modeling and Simulation: An Interdisciplinary Guide: An Interdisciplinary Guide*, vol. 21. New York: Springer New York.

Schneider, T., Lan, S., Stuart, A. M. & Teixeira, J. (2017) Earth system modeling 2.0: a blueprint for models that learn from observations and targeted high-resolution simulations. *Geophys. Res. Lett.*, **44**, 12–396.

Schneider, T., Stuart, A. M. & Wu, J.-L. (2020) Ensemble Kalman inversion for sparse learning of dynamical systems from time-averaged data. arXiv preprint arXiv:2007.06175.

Schütte, C. & Sarich, M. (2013) *Metastability and Markov State Models in Molecular Dynamics*, vol. 24. Providence, Rhode Island: American Mathematical Soc.

Selten, F. (1995) An efficient empirical description of large-scale atmospheric dynamics. *Ph.D. Thesis*. Turbulentie en Werveldynamica.

Sisson, S. A., Fan, Y. & Beaumont, M. (2018) *Handbook of Approximate Bayesian Computation*. Boca Raton, Florida: Chapman and Hall/CRC.

Smith, A. (2013) *Sequential Monte Carlo Methods in Practice*. New York: Springer Science & Business Media.

Tuckerman, M. (2010) *Statistical Mechanics: Theory and Molecular Simulation*. Oxford, England: Oxford University Press.

Tziperman, E., Stone, L., Cane, M. A. & Jarosh, H. (1994) El Niño chaos: overlapping of resonances between the seasonal cycle and the pacific ocean-atmosphere oscillator. *Science*, **264**, 72–74.

Tziperman, E., Cane, M. A., Zebiak, S. E., Xue, Y. & Blumenthal, B. (1998) Locking of El Niño's peak time to the end of the calendar year in the delayed oscillator picture of ENSO. *J. Climate*, **11**, 2191–2199.

Van Leeuwen, P. J. & Evensen, G. (1996) Data assimilation and inverse methods in terms of a probabilistic formulation. *Mon. Weather Rev.*, **124**, 2898–2913.

Wilkinson, D. J. (2018) *Stochastic Modelling for Systems Biology*, 3rd edn. Boca Raton, Florida: Chapman and Hall/CRC.

Wood, S. N. (2010) Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, **466**, 1102–1104.

Xiao, H., Wu, J.-L., Wang, J.-X., Sun, R. & Roy, C. (2016) Quantifying and reducing model-form uncertainties in Reynolds-averaged Navier–Stokes simulations: a data-driven, physics-informed Bayesian approach. *J. Comput. Phys.*, **324**, 115–136.

Ying, Y., Maddison, J. & Vanneste, J. (2019) Bayesian inference of ocean diffusivity from Lagrangian trajectory data. *Ocean Model.*, **140**, 101401.

Zhang, L., Mykland, P. A. & Aït-Sahalia, Y. (2005) A tale of two time scales. *J. Am. Stat. Assoc.*, **100**, 1394–1411.

Zhang, L., et al. (2006) Efficient estimation of stochastic volatility using noisy observations: a multi-scale approach. *Bernoulli*, **12**, 1019–1043.

Zhang, W., Hartmann, C. & Schütte, C. (2017) Effective dynamics along given reaction coordinates, and reaction rate theory. *Faraday Discuss.*, **195**, 365–394.